



Université  
de Toulouse

# THÈSE

En vue de l'obtention du

## DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par :

Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier)

Présentée et soutenue par :

**Eléna Cadic**

Le vendredi 14 février 2014

---

**Titre :**

Recherches de facteurs génétiques impliqués dans l'élaboration du rendement sous contrainte hydrique chez le tournesol *Helianthus annuus* par génétique d'association et analyse de liaison dans une population recombinante

---

École doctorale et discipline ou spécialité :

ED SEVAB : Interactions plantes-microorganismes

**Unité de recherche :**

Laboratoire des interactions plantes microorganismes

**Directeur(s) de Thèse :**

Patrick Vincourt

Brigitte Mangin

**Rapporteurs :**

Mathilde Causse

Christophe Plomion

**Autre(s) membre(s) du jury :**

Brigitte Crouau-Roy

Marie Coque

Felicity Vear



---

# Remerciements

Tout d'abord, je suis heureuse de pouvoir remercier Patrick Vincourt et Brigitte Mangin qui m'ont énormément apporté pendant ces années de thèse. Merci Patrick pour tes idées, ta disponibilité, ton soutien et bien sûr pour m'avoir accueilli dans ton équipe. Merci Brigitte pour ton temps. Je garderai en mémoire nos rendez-vous du mercredi où tu as pris le temps de m'expliquer des concepts et de m'aiguiller dans la bonne direction. Je vous remercie également d'avoir accepté de me suivre à distance dans la dernière ligne droite de la rédaction. Je remercie Pascual Perez et Bruno Grèzes-Besset pour m'avoir permis de faire cette thèse à Biogemma. Je remercie Marie Coque qui a repris le flambeau de mon encadrement au sein de Biogemma.

Merci aux membres du jury et en particulier aux rapporteurs : Mathilde Causse, Christophe Plomion ainsi qu'à Félicity Vear.

Je remercie les membres de mon comité de thèse dont Alain Charcosset et Philippe Debaeke pour avoir accepté de suivre l'évolution de la thèse.

Merci à mes collègues de Biogemma qui m'ont accueilli chaleureusement et avec qui j'ai partagé des moments inoubliables au boulot et en dehors...Isabelle, pour ta bonne humeur et les fous rires, Nicole, Pierre (pour les discussions phénotypage), Clotilde (pour la partie bioanalyse), Michèle (pour ton aide sur les données phéno), Jean-Luc, Nicolas (pour la cartographie), Fabienne, Nathalie, Delphine, Sabrina, Sylvie et bien sûr merci à Amandine pour avoir partagé son expérience de thésarde et avoir été à l'écoute de tous mes questionnements « d'avenir »...Merci pour vos contributions à ce travail de thèse et vos conseils. J'espère vous revoir encore très vite à Clermont.

Je remercie également mes collègues de l'INRA, la bande des thésards : Falah, Parham, Quentin et Gwennaëlle et avec qui j'ai grelotté dans le bureau...et les anciens Amandine, David et Yannick, qui ne sont plus dans l'équipe aujourd'hui mais pour qui j'ai une pensée. Merci aux Nicolas dont Nicolas Langlade (pour ton aide sur l'article et la partie ABA qui a été une expérience intéressante), Marie-Claude, Didier (avec qui on s'est bien appliqué à préparer les sushis en chambre de culture), Stéphane (qui a réussi à me convaincre de courir 10km....)



Je souhaite également remercier Pierre Casadebaig et Philippe Debaeke pour leurs conseils sur le modèle SUNFLO.

Merci aux collègues d'Euralis et de Soltis, qui ont contribué à ce travail, notamment pour la partie phénotypage et cartographie, Philippe Blanchard, Marie Boillot ainsi que Thierry André. Merci aux partenaires du projet, les collègues de Syngenta et RAGT qui ont participé à la mise en place du réseau d'essais.

Merci à Eliette Combes qui a accepté de me confier un nouveau poste malgré que la rédaction de thèse n'était pas finie, merci à mes nouveaux collègues de Limagrain pour m'avoir soutenu dans la dernière ligne droite.

Enfin, merci à ma famille et mes amis qui, je crois avaient hâte que tout ça se finisse...promis je reviendrai plus souvent en Bretagne !! Un énorme merci à Adrien pour son soutien sans faille, tu as été d'une grande aide pour moi, je mesure chaque jour la chance que j'ai. Un grand merci à tous! Je vais enfin finir de déballer les cartons....



---

# Abstract

For most crops, water availability is a major component of yield. These resources, already scarce today, will be increasingly scarce in the future. We must improve crop plants tolerance to water deficit in order to guarantee food security. Sunflower, a species of economic importance is also concerned by this challenge. The aim of this thesis is to identify genomic regions involved in the variability of yield and its components under drought. To this purpose, an association mapping approach has been led on a core collection of 384 cultivated and elite lines collected from different seeds companies. This panel has been evaluated as hybrids combinations with different testers in a multi environment trial of 17 environments for phenology, productivity and senescence traits. In a first step, the separated analysis per environment led to the identification of 157 markers associated with at least one trait over a total of 6000 SNP tested by apply a model taking into account the structuration in two groups: male restorer lines and female lines, and the relatedness between lines. Among these significant markers, 34 were associated with productivity but only on a specific environment, underlying the importance of genotype by environment interaction. In a second step, a crop model (SUNFLO) simulating sunflower genotypes yield depending on environmental conditions, led to the characterization of drought stress inside the multi environment trial. By using varieties defined before in the model, a drought index has been estimated for each environment. Thus, the panel response to this index has been tested in association mapping leading to the identification of several new locus involved in drought stress response and yield stability. Finally, in a complementary approach, a bi parental population of 273 recombinant inbred lines has been evaluated on 6 environments for most traits or 16 environments for flowering date. QTL mapping on this population confirmed some regions and led to the detection of new ones.

Keywords: drought stress, yield, QTL, association mapping





---

# Résumé

La disponibilité en eau est un facteur essentiel pour le rendement des principales espèces cultivées. Dans un contexte où les ressources sont déjà rares et risquent de le devenir davantage, il est aujourd'hui impératif d'améliorer la tolérance des plantes au stress hydrique afin de garantir la sécurité alimentaire. Le tournesol figurant parmi les espèces d'importance économique majeure est également concerné par cet enjeu. L'objectif de cette thèse est d'identifier les régions génomiques permettant d'expliquer la variabilité du rendement et de ses composantes sous contraintes hydriques. Pour cela, une approche de génétique d'association a été menée sur un panel de 384 lignées cultivées et élites provenant de différentes entreprises semencières. Ce panel a été évalué en combinaison hybride avec plusieurs testeurs sur un réseau expérimental de 17 environnements et pour des caractères de phénologie, de sénescence et de productivité. Dans un premier temps, l'analyse séparée de chaque environnement a permis d'identifier 157 marqueurs associés avec au moins un caractère sur un total de 6000 SNP testés grâce à l'utilisation d'un modèle prenant en compte la structuration du panel en deux groupes : les lignées mâles (restauratrices de fertilité) et les lignées femelles (mainteneuses de stérilité mâle), ainsi que leur niveau d'apparentement. Parmi ces marqueurs significatifs, 34 étaient associés avec un caractère lié à la productivité, mais ceci, le plus souvent, de manière spécifique à un environnement, soulignant ainsi l'importance de l'interaction génotype-environnement. Dans un second temps, un modèle éco-physiologique (SUNFLO) simulant la performance de génotypes de tournesol en fonction des conditions environnementales, a permis de caractériser le stress hydrique au sein du réseau expérimental. En utilisant des variétés témoins préalablement définies dans le modèle, un index de stress a été estimé pour chaque environnement. La réponse du panel à cet index a ensuite été testée en génétique d'association permettant d'identifier de nouveaux locus impliqués dans la tolérance au stress hydrique et la stabilité du rendement.

Enfin, en complément de cette approche de génétique d'association, une population biparentale de 273 lignées recombinantes (RIL) a été évaluée sur 6 environnements pour la plupart des caractères ou sur 16 environnements pour la floraison. La détection de QTL à partir de cette population a permis de confirmer certaines zones détectées en génétique d'association et d'en découvrir de nouvelles.

Mots clés : stress hydrique, rendement, QTL, génétique d'association



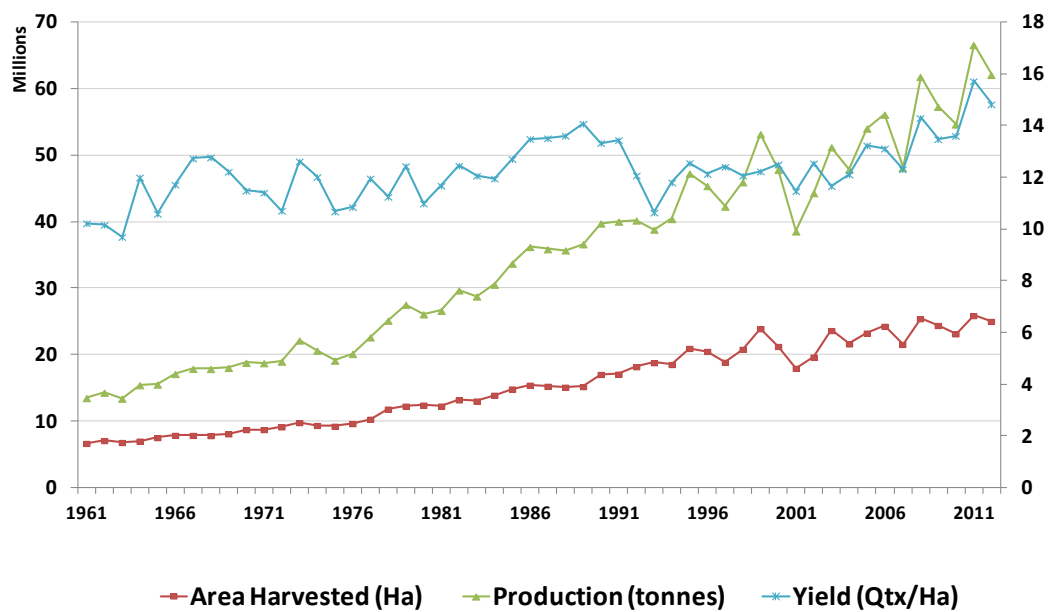
---

# Tables des matières

<b>INTRODUCTION GENERALE.....</b>	<b>8</b>
<b>Chapitre I Amélioration génétique de la productivité dans un contexte de renforcement des interactions génotype - environnement engendré par l'aléa de la disponibilité en eau : synthèse bibliographique.....</b>	<b>16</b>
I.1 Contexte et enjeux.....	16
I.2 Quel type de « contrainte hydrique » prendre en compte ?.....	17
I.3 Stratégies d'adaptation des plantes à la sécheresse.....	18
I.4 Mécanismes moléculaires impliqués dans la réponse au stress.....	19
I.5 Résumé des caractères contrôlant la réponse des plantes au stress hydrique.....	22
I.6 Quels sont les progrès obtenus pour la tolérance à la sécheresse en utilisant la sélection assistée par marqueurs?.....	27
I.7 Stratégies d'analyses de l'interaction GE.....	30
<b>Chapitre II Analyse des données phénotypiques.....</b>	<b>34</b>
II.1 Introduction.....	34
II.2 Analyse des données phénotypiques par environnement.....	34
II.2.1 Matériels et méthodes.....	34
II.2.1.1 Matériel végétal.....	34
II.2.1.2 Dispositif et caractères mesurés.....	35
II.2.1.3 Analyses statistiques.....	37
II.2.2 Résultats et discussions.....	39
II.2.2.1 Ajustement des données phénotypiques avec le modèle naïf.....	39
II.2.2.2 Ajustement des données phénotypiques avec les modèles spatiaux.....	41
II.2.2.3 Intérêt des caractères choisis.....	41
II.3 Analyse multilocale.....	45
II.3.1 Interaction génotype x environnement.....	45
II.3.1.1 Matériels et méthodes.....	45
II.3.1.2 Résultats et discussion.....	46
II.3.2 Description du modèle SUNFLO.....	49
II.3.3 Caractérisation des environnements avec le modèle SUNFLO.....	51
II.3.3.1 Matériels et méthodes.....	51
II.3.3.2 Résultats et discussion.....	54
II.3.4 Mise en place d'un index synthétique de réponse au stress pour le panel d'association... 57	
II.3.4.1 Matériels et méthodes.....	57
II.3.4.2 Résultats et discussion.....	59
<b>Chapitre III Analyse du panel et choix de modèles de génétique d'association.....</b>	<b>62</b>
III.1 Etude bibliographique : la génétique d'association chez les plantes.....	62
III.2 Panel : origine et données moléculaires.....	66
III.2.1 Origine du matériel.....	66
III.2.2 Données moléculaires.....	67
III.3 Structuration du panel.....	68
III.3.1 Matériels et méthodes.....	68
III.3.2 Résultats et discussion.....	70
III.3.2.1 Inférence de la structure à partir des SNP.....	70



III.3.2.2	Comparaison avec les résultats obtenus à partir des SSR.....	71
III.4	Déséquilibre de liaison .....	73
III.4.1	Matériels et Méthodes .....	73
III.4.2	Résultats et discussion.....	74
III.5	Comparaison des modèles de génétique d'association.....	76
III.5.1	Matériels et méthodes.....	77
III.5.2	Résultats et discussion.....	78
<b>Chapitre IV</b>	<b>Locus impliqués dans le déterminisme génétique du rendement sous contraintes hydriques.....</b>	<b>80</b>
IV.1	Introduction .....	80
IV.2	Matériels et méthodes.....	81
IV.3	Résultats et discussion.....	83
IV.3.1	Résultats des tests d'association sur chaque combinaison environnement-caractère....	83
IV.3.1.1	Test multiples .....	83
IV.3.1.2	Comparaison des modèles Kais et Kais + Testeur .....	84
IV.3.1.3	Résultats du modèle Kais + Testeur selon les environnements.....	85
IV.3.1.4	Résultats du modèle Kais + Testeur selon les caractères .....	86
IV.3.1.5	Zones génomiques détectées .....	88
IV.3.2	Résultats et discussion des associations sur différents index de réponse au stress au travers du réseau multilocal.....	90
IV.3.2.1	Introduction .....	90
IV.3.2.2	Résultats des analyses d'associations sur les variables synthétiques .....	92
IV.4	Synthèse des régions d'intérêt.....	94
IV.4.1	Introduction .....	94
IV.4.2	Résultats .....	94
<b>Chapitre V</b>	<b>Détection de QTL dans une population biparentale et comparaison avec l'approche « génétique d'association » .....</b>	<b>98</b>
V.1	Introduction .....	98
V.2	Etude sur le caractère floraison .....	101
V.2.1	Résumé de l'article .....	101
V.2.2	Article.....	103
V.3	Etude sur les autres caractères agronomiques .....	104
V.3.1	Matériels et méthodes.....	104
V.3.2	Résultats et discussion.....	105
V.3.2.1	Analyses phénotypiques .....	105
V.3.2.2	Détection de QTL.....	106
V.3.2.3	Colocalisations entre les associations et QTL bi parentaux .....	108
<b>DISCUSSION.....</b>		<b>110</b>
<b>CONCLUSION.....</b>		<b>116</b>
<b>REFERENCES .....</b>		<b>117</b>
<b>ANNEXES.....</b>		<b>119</b>



**Figure 1 : Evolution des surfaces, de la production et du rendement du tournesol.**  
 (Source : FAOSTAT, septembre 2013)

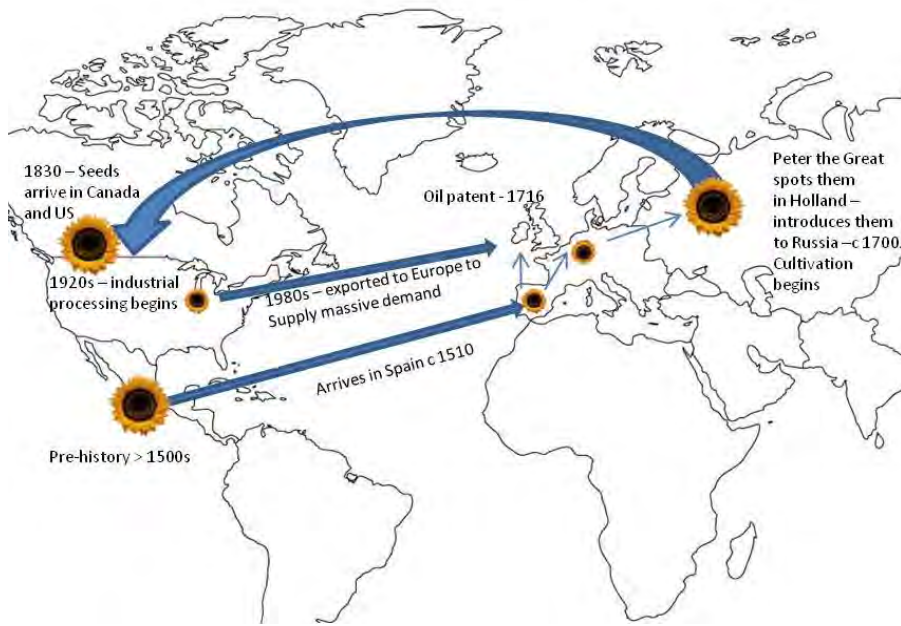
---

# Introduction générale

## *Importance économique*

Le tournesol est la quatrième culture oléagineuse mondiale avec 25 millions d'hectares cultivés en 2012 (FAOSTAT). La production a augmentée de 40 % en 20 ans pour atteindre aujourd'hui 37 millions de tonnes (Figure 1). Les principaux producteurs sont la Russie et l'Ukraine, qui détiennent quasiment la moitié de la production mondiale puis l'UE (Roumanie, Bulgarie, Espagne, France) et l'Argentine. Alors que les surfaces ont diminué en Argentine et aux Etats-Unis notamment en raison de la concurrence du soja (Jouffret *et al.*, 2011), on observe une forte augmentation des surfaces dans les pays de l'Est de l'Europe (UE ainsi que Russie et Ukraine). Ces nouvelles terres dédiées au tournesol et amenées à s'étendre, constituent un enjeu réel pour la France, premier producteur européen et deuxième exportateur mondial de semences.

Le tournesol est devenu une espèce stratégique aux nombreux débouchés. L'huile pour l'alimentation humaine en est le principal débouché et représente 57% du marché des huiles de table en France. Riche en acide gras insaturés, en vitamine E aux propriétés anti-oxydantes et en phytostérols, dont les effets anti-cholestérol sont reconnus, l'huile de tournesol bénéficie d'une image favorable du point de vue de la santé. Le tourteau pour l'alimentation animale, coproduit de l'huilerie (55% de la graine), constitue la deuxième utilisation des graines de tournesol. Encouragée pour diminuer la dépendance face aux importations de soja américain, l'incorporation de tourteau de tournesol dans les rations animales est restée longtemps minoritaire du fait de la présence des coques. Elle est actuellement en cours de progression grâce à la mise en œuvre de nouvelles technologies de décorticage. Moins de 10% de la production annuelle de tournesol est destinée à des usages non alimentaires. Le secteur de l'oléochimie permet de valoriser l'huile de tournesol sous forme de glycérine végétale dans la cosmétique, la pharmacie; comme biolubrifiants dans l'industrie automobile, la métallurgie; ou encore comme solvants non toxiques, revêtements et encres. L'arrivée sur le marché, ces dernières années, de variétés oléiques contenant de 60 à 80% d'acide oléique (contre 15 à 20% pour le tournesol classique), offre de nouveaux débouchés, alimentaires: huiles recombinaisons du type ISIO 4 et industriels : base oléochimique pour lubrifiant, biopolymères et biodiesel, composé de 2 à 5% d'huile de tournesol. Avec 56% des surfaces de tournesol consacrées aux variétés oléiques en 2011, la France est leader sur ce marché.



**Figure 2 : Histoire de l'évolution du tournesol (<http://www.kuriositas.com/2011/08/strange-history-of-sunflower.html>)**



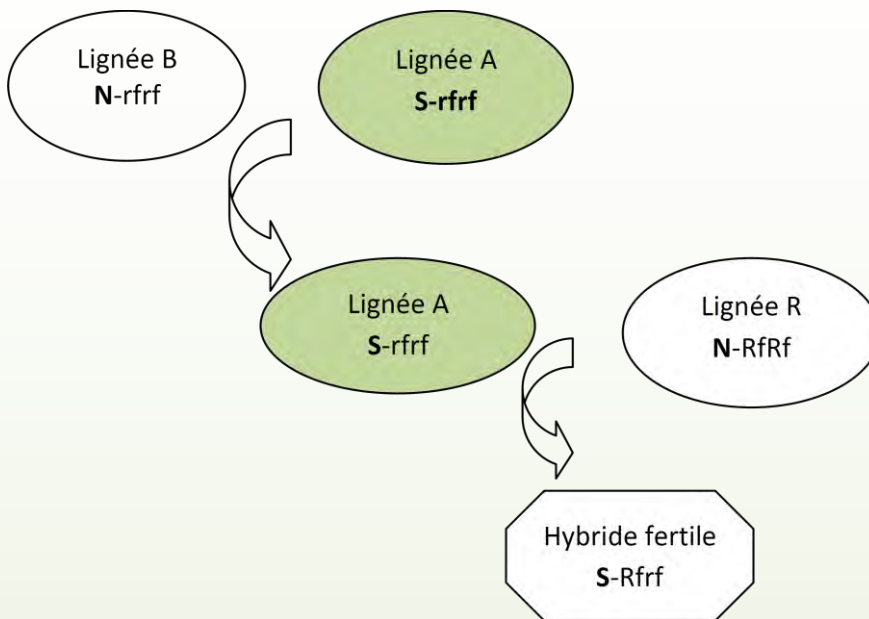
Une augmentation de la demande en huile végétale est attendue à l'horizon 2015-2020, avec la croissance de la Chine et de l'Inde. La qualité nutritionnelle de l'huile de tournesol associée à la durabilité de sa production en fait une culture attrayante pour répondre aux besoins tout en prenant en compte les contraintes environnementales. En effet, le tournesol demande peu d'intrants azotés ou de traitements phytosanitaires.

### ***Histoire de l'amélioration du tournesol***

Le tournesol provient d'Amérique du nord, où il aurait été domestiqué vers 4300 ans av. JC. (Heiser *et al.*, 1969, Blackman *et al.*, 2011 ) (Figure 2). Les Indiens utilisaient alors les graines de tournesol dans leur alimentation mais aussi pour ses vertus médicinales, ou encore comme teinture pour les textiles. Les explorateurs espagnols ont ensuite apporté le tournesol en Europe en 1510 où il était principalement ornemental. Ce n'est qu'à partir du 19<sup>ième</sup> siècle que les russes ont cultivé le tournesol à grande échelle pour son huile, l'une des seules huiles alimentaires autorisées par l'église orthodoxe durant le carême. Avec la commercialisation du tournesol en 1830 se sont développés les premiers efforts de sélection pour améliorer son contenu en huile. La sélection massale, mise en place dans la station de Krasnodar en Russie a permis d'obtenir des résultats sur la maturité, la teneur en huile et la résistance aux maladies et aux insectes. (Morozove *et al.*, 1947 ; Vranceanu *et al.*, 1974 ). Ce sont les premiers programmes de sélection récurrente mise en place par Pustovoit (1967) qui ont permis des améliorations de la teneur en huile très significatives. La méthode qu'il développa, appelée « méthode des semences restantes » est basée sur l'étude des descendance et la création de nouvelles populations à partir du reste des semences des meilleurs individus (Gallais, 1992). Les teneurs en huile dans les grains ont ainsi augmenté de 330g/Kg à 550g/Kg entre 1940 et 1964 (Pustovoit, 1964). Finalement, une partie des ressources génétiques constituant le fondement des variétés actuelles ont été rapportées en Amérique du nord et en particulier au Canada par les immigrants russes à la fin du 19<sup>ième</sup> siècle.

Alors que les variétés populations dominaient le marché, la découverte en France de systèmes de stérilité male génique puis cytoplasmique (CMS) (Leclercq, 1969) à partir de l'espèce *Helianthus petiolaris* (CMS-PET1), combinée à celle des gènes de restauration de la fertilité (Kinman, 1970) a popularisé la culture du tournesol hybride en Europe et à travers le monde. Les premiers hybrides cytoplasmiques ont été cultivés à partir de 1978. Aujourd'hui le

## Production des hybrides commerciaux de tournesol



**N** : cytoplasme normal

**S** : cytoplasme stérilisant

**Rf** : allèle dominant pour la restauration de la fertilité

**rf** : allèle récessif dont la présence à l'état homozygote entraîne le maintien de la stérilité mâle

Les lignées de type A possèdent un cytoplasme stérilisant (S) et sont homozygotes pour le gène de maintien de la stérilité mâle. Elles sont donc mâles stériles : les anthères sont petites et aucun pollen n'est visible à la floraison. Ces lignées A constituent donc les parents femelles des hybrides commerciaux mais comme le cytoplasme stérilisant se transmet par l'ovule, il est nécessaire que le parent mâle qui apporte le pollen puisse annuler cet effet stérilisant de manière à obtenir des hybrides fertiles, avec la production de graines attendue. Les lignées de type R, appelés restaurateurs de la fertilité (R), le permettent. Ce sont en effet des lignées homozygotes pour l'allèle dominant du gène de restauration de la fertilité.

La lignée A parent de l'hybride est obtenu par rétrocroisement d'une lignée d'intérêt, que l'on appelle lignée B, avec une lignée A au cytoplasme stérile. La lignée A obtenue et la lignée B initiale sont donc quasi-isogéniques. Les lignées de type B sont appelées « mainteneuses de stérilité » car elles sont homozygotes pour l'allèle récessif rf.

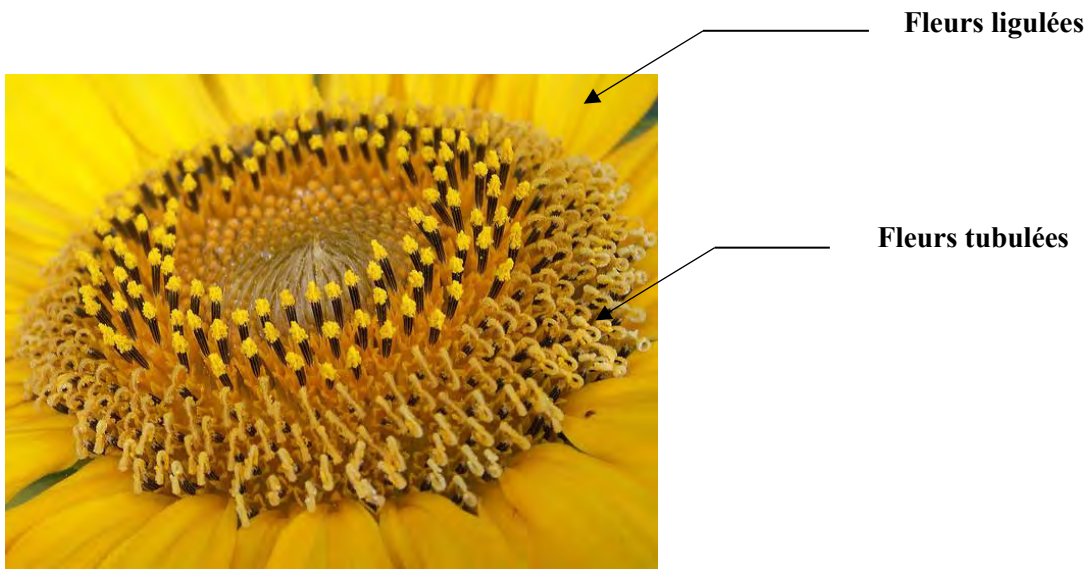
tourneol est le deuxième hybride cultivé après le maïs et le système CMS-PET1 est le principal utilisé pour la production des semences hybrides (cf. encadré).

Comme pour d'autres espèces allogames, l'hétérosis est significatif chez le tourneol, notamment pour le rendement et la hauteur (Duvick D.N, 1997, Cheres, 2000). Contrairement au maïs, chez le tourneol les groupes hétérotiques n'ont pas été clairement déterminés (Cheres *et al.*, 2000) et même si la plupart des études visant à structurer la diversité à partir des marqueurs moléculaires séparent les lignées mainteneuses de stérilité des lignées restauratrices, (Tersac *et al.*, 1993, 1994 ; Gentzbittel *et al.*, 1994 ; Zhang *et al.*, 1995 ; Hongtrakul *et al.*, 1997 ; Mandel *et al.*, 2011), cette séparation résulte davantage de l'histoire qui a favorisé la sélection de chacun des groupes indépendamment plutôt que de la recherche de l'aptitude à la combinaison. En dehors de l'hétérosis, la création variétale centrée sur la production d'hybrides a permis de répondre rapidement aux problèmes posés par les bioagresseurs lors de l'extension du tourneol sur nos territoires (Vincourt et Vear, 2009) en apportant des lignées plus diversifiées ou intégrant du patrimoine d'espèces sauvages apparentées intéressantes dans les combinaisons hybrides.

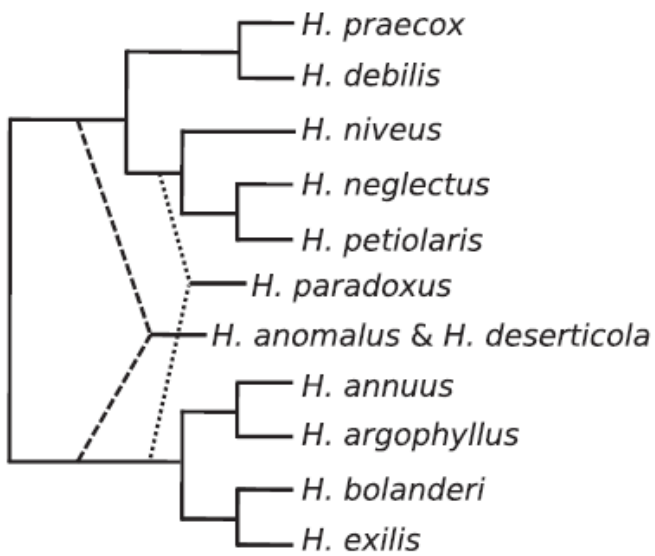
### ***Biologie et évolution des génomes***

Le tourneol (*Helianthus annuus*) appartient à la famille des composées et possède une inflorescence de type capitule, caractéristique de cette famille, comportant de 300 à 1000 fleurs. Les fleurs de l'extérieur du capitule sont ligulées, stériles et portent trois pétales jaunes soudés (Figure 3). Les fleurs du centre ou fleurons sont tubulées, fertiles et hermaphrodites. Le tourneol est une plante entomophile pour laquelle la présence d'insectes pollinisateurs est essentielle durant la floraison. Ils rendraient même la maturité des graines plus homogène. Le fruit du tourneol est appelé akène et est muni d'un péricarpe membraneux (la coque), de couleur blanchâtre à noirâtre, non soudé à la graine et représentant de 18 à 40% du poids du fruit. Il entoure une amande contenant 55 à 70% d'huile.

Le tourneol sauvage dont est issu le tourneol cultivé est caractérisé par la présence d'une ramification et donc de plusieurs capitules aux akènes de faible taille dispersés à maturité (Burke et al, 2005) ; c'est un allogame obligatoire tandis que le tourneol cultivé est monocapité, avec des akènes plus gros et a perdu son auto-incompatibilité sporophytique (Liu *et al.*, 2006). Suite au goulot d'étranglement subi lors de la domestication, le polymorphisme génétique du tourneol cultivé a été réduit de 40 à 50% en comparaison de la diversité observé



**Figure 3 : Détails de l'inflorescence du tournesol**



**Figure 4: Arbre phylogénétique du tournesol commun inféré à partir d'analyses d'ADN ribosomal et de 11 gènes simple copies (Kane et al., 2012)**

dans le compartiment sauvage (Micic *et al.*, 2005). Cependant, les tournesols sauvages et domestiqués appartiennent à la même espèce et sont donc interfertiles (Snow, 1998), ce qui a permis au tournesol cultivé de bénéficier de nombreuses introgressions d'allèles sauvages mais aussi provenant d'autres espèces du genre *Helianthus*.

Le genre *Helianthus* est riche de 51 espèces annuelles ou pérennes. La figure 4 présente la taxonomie des espèces annuelles du genre *Helianthus*. L'espèce la plus proche du tournesol commun *Helianthus annuus* est l'espèce sauvage *Helianthus petiolaris* (à l'origine du système de stérilité mâle cytoplasmique). On distingue deux clades sur la figure 4 avec des espèces intermédiaires dérivant d'hybridations entre ces deux clades (Kane *et al.*, 2012). Le genre *Helianthus* possède ainsi une histoire riche d'hybridations qui ont abouti à des espèces aux génomes très variés.

Si beaucoup de caractères des espèces sauvages ne sont pas profitables pour le rendement, certains comme la résistance aux maladies ont permis de nombreuses avancées, grâce à la coévolution du tournesol avec les pathogènes. Par exemple, les gènes de résistance au mildiou semblent assez fréquents dans les formes sauvages d'*H. annuus* et les autres espèces annuelles (*H. argophyllus* notamment) (Vear *et al.*, 2008). Les premières lignées résistantes au mildiou ont d'ailleurs été obtenues à partir d'une autre espèce *H. tuberosus* (topinambour).

Les hybridations du tournesol cultivé avec le tournesol sauvage de la même espèce ou d'autres espèces ont permis au tournesol de coloniser de nombreux habitats et de s'adapter aux contraintes abiotiques, telles que la résistance à la sécheresse (Whitney *et al.*, 2006). On observe des espèces du genre *Helianthus* du Mexique au Canada, adapté à des milieux très différents, tels que des milieux arides (*H. deserticola*), des dunes (*H. anomalus*) ou encore des milieux salins (*H. paradoxus*). Ces hybridations qui ont eu de nombreuses conséquences sur l'évolution des génomes, le tournesol ayant d'ailleurs le taux d'évolution kariotypique le plus élevé dans l'ensemble des règnes animaux et végétaux (Burke *et al.*, 2004), font du tournesol un modèle d'étude riche pour la spéciation (Rieseberg, 1995).

### ***Cibles d'amélioration***

Depuis 30 ans, l'amélioration génétique du tournesol a permis un gain de rendement de 0.5 quintaux/ha/an. En situations favorables, le rendement moyen en grande parcelle peut atteindre 45 quintaux/ha. En Russie et en Ukraine, les rendements connaissent de fortes fluctuations liées aux conditions climatiques. De plus, la production y est encore assez extensive et encore largement élaborée à partir de variétés populations ou d'hybrides locaux



développés récemment à partir de matériel peu performant. En France, le rendement progresse lentement comparé au progrès estimé dans les essais dédiés à l'inscription des variétés au Catalogue Officiel, ceci étant dû au fait que ces essais sont conduits en général dans des conditions de culture plus favorables que l'aire de culture du tournesol qui s'est orientée vers des terres de moindre potentiel. Dans un contexte de réglementations renforcées sur l'utilisation des intrants et en conjonction avec le changement climatique, il est nécessaire de continuer à améliorer par la voie génétique le rendement du tournesol et à le stabiliser dans des conditions environnementales fluctuantes.

Pour cela, les recherches publique et privée se sont organisées autour de ces enjeux majeurs. La France occupe une place dominante dans l'amélioration génétique du tournesol. En effet, le marché se développant dans les pays de l'Est de l'Europe, le tournesol est devenu une espèce hautement stratégique pour les entreprises semencières. Celles-ci sont très largement implantées dans le sud-ouest, la région Midi-Pyrénées étant la première région de production de tournesol en France (estimation 2013 de la surface totale : 730000 ha). C'est à l'Institut National de la Recherche Agronomique de Toulouse que se trouve le pôle « Agrogénomique du Tournesol », qui s'appuie principalement sur deux trois unités de recherche : le Laboratoire des Interactions Plantes Microorganismes (LIPM), au sein duquel j'ai effectué mon travail de thèse, l'UMR AGIR (affiliée notamment au Département INRA « Environnement et Agronomie ») et le Centre National de Ressources Génomiques Végétales (CNRGV). Dans le cadre de la recherche sur le tournesol, le LIPM est centré à la fois sur les problématiques biotiques telles que celles provenant du mildiou et du phoma, deux pathogènes extrêmement dommageables pour le tournesol, ainsi que sur des problématiques abiotiques, comme la tolérance à la sécheresse. Le CNRGV, centre national des ressources génétiques végétales est impliqué notamment dans le séquençage du génome du tournesol et l'UMR AGIR travaille sur les approches agronomiques et de modélisation. Témoignant de l'importance de la recherche sur l'amélioration du tournesol, depuis 2006 se succèdent plusieurs projets visant à améliorer l'état des connaissances sur cette espèce dont notamment les projets SUNYFUEL (ANR) et OLEOSOL (soutenu par le Fonds Interministériel de soutien aux Pôles de Compétitivité : FUI), en partenariat avec les industries semencières, et dans lesquels s'inscrit cette thèse. Plus récemment, le projet Investissement d'avenir SUNRISE financé à hauteur de 7 millions d'euros sur 8 ans (<http://www.sunrise-project.fr/>) a été lancé en 2012 avec comme objectif d'améliorer la production d'huile du tournesol en conditions hydriques limitées, par des approches génétiques, génomiques et écophysologiques.





## ***Outils génétiques et génomiques pour l'amélioration du tournesol***

De nombreuses ressources génétiques sont disponibles chez le tournesol. Des collections sont maintenues et caractérisées dans des centres de ressources génétiques notamment aux Etats-Unis (USDA sunflower genebank, Iowa : 2797 accessions cultivées et 1344 sauvages d'*H. annuus*) et en France (INRA Toulouse). Outre ces accessions sauvages, la collection française comprend des variétés populations (environ 400), des lignées de toute origine (Europe, Russie, Canada, Argentine, Afrique, Asie,...) au nombre de 2200 (Vear *et al.*, 2011). Les premières études de structuration de la diversité ont été réalisées essentiellement sur la base des données phénotypiques pour la recherche de résistance au mildiou, au sclerotinia, puis au phomopsis. Puis l'accumulation de données génotypiques grâce à l'utilisation des marqueurs moléculaires a permis de préciser la structuration et d'élaborer des « core collections » capables de synthétiser au mieux la diversité existante. Toutes ces ressources offrent ainsi un potentiel très important pour l'amélioration du tournesol. Pour répondre aux questions liées à la domestication et à l'évolution des génomes, des populations de cartographies entre parents sauvages et domestiqués, entre lignées élites ou locales (landraces) ont été créés (Burke *et al.*, 2002, Wills and Burke, 2007 ; Baack *et al.*, 2008 ; Bowers *et al.*, 2012). D'autres méthodes pour créer de la variabilité génétique ont été explorées telles que la mutagenèse chimique qui permet d'obtenir des populations de mutants dont les mutations peuvent ensuite être criblées par PCR. Cette méthode appelée TILLING (Targeting Induced Local Lesion IN Genome) a été récemment mise en place chez le tournesol (Sabetta *et al.*, 2011 ; Kumar *et al.*, 2013).

Grâce à la diminution des coûts des plateformes de séquençage et de génotypage, on assiste ces dernières années au développement massif des ressources génomiques pour le tournesol. Des cartes génétiques à haute densité sont maintenant disponibles. Bowers *et al.*, (2012) ont utilisé quatre populations développées à partir de parents d'origine diverses (lignées élites, sauvages ou locales, sélectionnées pour la confiserie ou pour l'huile), génotypées sur la plateforme iScan d'Illumina (Illumina Inc., San Diego, CA) pour créer une carte consensus de 10 000 locus. Le développement de ces cartes à haute densité est non seulement utile à la dissection de caractères d'intérêts mais facilite également l'assemblage du génome (Kane *et al.*, 2011).

Le tournesol ( $2n=34$  chromosomes) possède un génome grand et complexe comparé à d'autres espèces, avec 3.5 Gb et de nombreux éléments répétés - principalement des rétrotransposons de type LRT (Long Terminal Repeat) - du fait notamment de son histoire jonchée d'évènements de polyploïdisation. Le séquençage du génome du tournesol cultivé a été entrepris en 2009 dans le cadre d'un consortium international dans lequel participent les



équipes du pôle agrogénomique de Toulouse. Des assemblages de bonne qualité mais qui couvrent essentiellement la partie génique sont d'ores et déjà disponibles. Cependant, jusqu'à présent, les bases de séquences EST ont constitué le point d'entrée majeur de la génomique pour des espèces non modèles comme le tournesol, notamment grâce à leur coût peu élevé (Lai *et al.*, 2012). Aujourd'hui plus de 400 000 EST sont disponibles dans les bases de données du NCBI. Ces EST ont pu être utilisés pour la découverte de SNP et aussi pour constituer un transcriptome de référence. Ainsi en assemblant les EST de sept espèces du genre *Helianthus*, une puce Affymetrix de 2.3 millions de sondes a pu être créée et utilisée pour analyser notamment les réponses différentielles à la sécheresse des transcriptomes de divers génotypes de tournesol (Rengel *et al.*, 2012).

### ***Conclusion***

Le tournesol est aujourd'hui une espèce stratégique mais dont la compétitivité est menacée par des rendements sous-optimaux. La stabilité du rendement à travers des contraintes abiotiques fluctuantes fait aujourd'hui parti des cibles majeures des programmes de sélection, d'autant que le réchauffement climatique risque d'amener plus fréquemment des situations de faible disponibilité en eau. Grâce à son importance économique et aussi au fait que le tournesol soit devenu une espèce modèle pour de nombreuses problématiques liées à l'évolution des génomes, la spéciation ou l'écologie, les ressources génomiques se sont développées massivement ces dernières années, la séquence complète du génome étant très bientôt disponible. De nombreux outils s'offrent donc aux généticiens et aux sélectionneurs pour comprendre le déterminisme des caractères complexes et améliorer la performance des hybrides. Parmi ces outils, la génétique d'association est apparue récemment comme une méthode de choix pour identifier les locus responsables de la variation des caractères phénotypiques car elle ne nécessite pas la création de populations dédiées ; la résolution avec laquelle sont cartographiés les locus d'intérêt est plus élevée que par le passé et elle permet d'exploiter une diversité plus large. La faible étendue du déséquilibre de liaison chez le tournesol (Kolman *et al.*, 2007) et la diversité présente, notamment dans le pool sauvage, font de la génétique d'association une méthode attractive pour l'étude du déterminisme génétique des caractères complexes chez le tournesol.



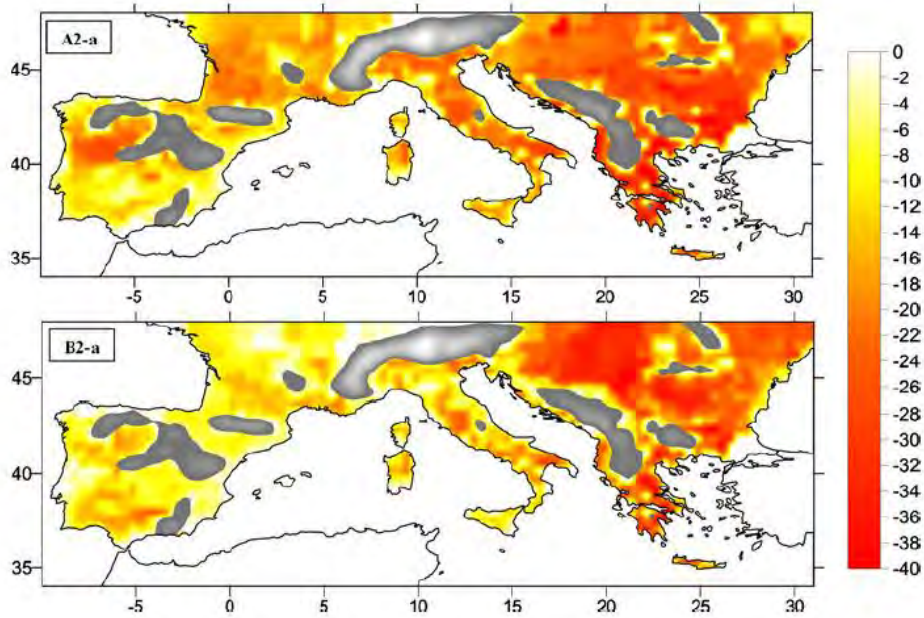
C'est à partir d'une core collection de lignées d'*Helianthus annuus* cultivées dont certaines font partie de collections de matériel public ou développé par l'INRA, d'autres sont « élites » - c'est-à-dire qu'elles sont impliquées dans la production d'hybrides commercialisés et largement cultivés - et d'autres enfin dérivent de lignées élites par introgression du patrimoine de différentes accessions de type sauvage, que nous avons appliqué, pour la première fois sur le tournesol, la génétique d'association afin d'identifier les locus impliqués dans le rendement sous contraintes hydriques.

Les objectifs de cette thèse sont :

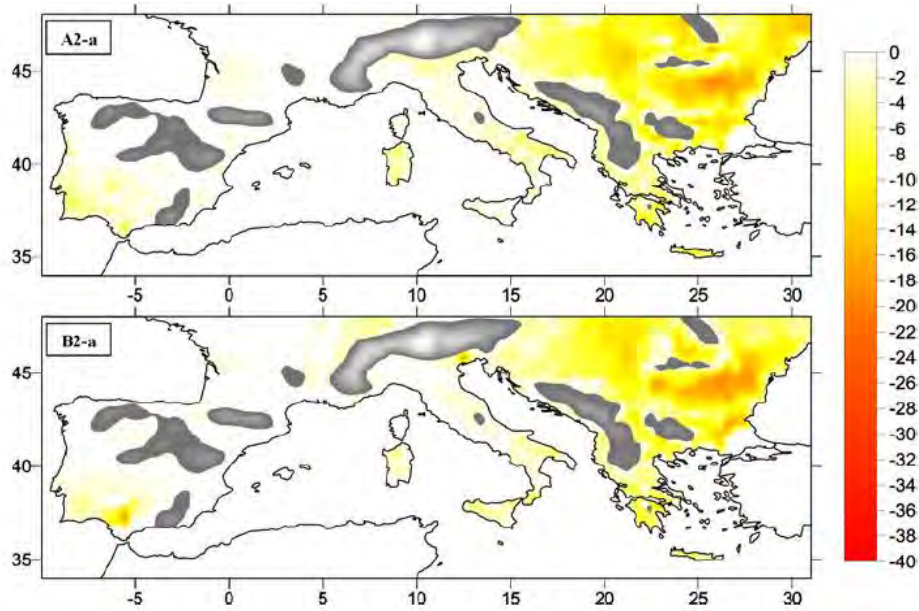
- Etudier la faisabilité de la génétique d'association sur le tournesol à partir d'un panel de lignées cultivées d'intérêt pour les sélectionneurs.
- Identifier des régions génomiques responsables de la stabilité du rendement sous contraintes hydriques.
- Comparer deux méthodologies de détection de QTL (« analyse de liaison » et génétique d'association).

Le chapitre I présente une revue bibliographique centrée sur la prise en compte de la tolérance à la sécheresse comme objectif de sélection dans le contexte des interactions génotype-environnement. Le chapitre II présente l'étude de la structuration du panel et de toutes les étapes préalables aux tests de génétique d'association. Le chapitre III traite de l'analyse des données phénotypiques par environnement et de la mise en place d'index synthétique de la tolérance des génotypes au stress hydrique. Le chapitre IV détaille les résultats des tests d'associations et le chapitre V décrit la comparaison de la génétique d'association avec une détection de QTL à partir d'une population biparentale pour le caractère « précocité de floraison » (sur la base d'un article publié) et les autres caractères agronomique d'intérêt.

Ce travail de thèse a été réalisé dans le cadre d'un contrat CIFRE financé par Biogemma et l'Agence Nationale de la Recherche. Les données ont été acquises grâce au programme HP1 de Génoplante, au projet ANR SUNYFUEL et au projet OLEOSOL, financé par la région Midi-Pyrénées, le Fonds Interministériel de soutien aux Pôles de Compétitivité (FUI), le Fonds Européen de Développement Régional (FEDER), le Conseil Général de l'Aveyron et la Communauté d'agglomération du Grand Rodez.



a



b

**Figure I.1 : Changement moyen de l'augmentation journalière de l'indice de récolte (DIR/dt (%)) dû à l'impact du stress thermique sur l'anthèse pour la période 2071-2100 par rapport à la période 1961-1990 (Moriondo et al. 2011).**

Scénario A2 : émissions de gaz à effets de serre moyennes à hautes, scénario B2 : émissions de gaz à effets de serre basses à moyennes ; a : tournesol, b: blé. Les zones grisées correspondent à des régions non cultivées pour ces 2 espèces.

# **Chapitre I Amélioration génétique de la productivité dans un contexte de renforcement des interactions génotype - environnement engendré par l'aléa de la disponibilité en eau : synthèse bibliographique.**

## **I.1 Contexte et enjeux**

La production agricole est directement affectée par les changements climatiques amenés à s'intensifier dans les années à venir (IPCC, 2007). En particulier les effets observés sont liés à la disponibilité en eau et aux augmentations de températures qui entraînent d'ores et déjà des baisses de rendement. Il a été montré à partir de la confrontation de données climatiques et de rendement pour la période 1980-2008, qu'une augmentation de 1°C s'accompagnait d'une baisse de 10% de rendement pour un panel de quatre espèces (maïs, blé, riz et soja) sur la plupart des zones cultivées (excepté pour le riz dont le rendement profite d'une hausse de température sur les hautes latitudes) (Lobell *et al.*, 2011). Au-delà des changements moyens de températures, la fréquence d'épisodes climatiques extrêmes risque d'augmenter, notamment lors des phases sensibles du cycle de culture telles que l'anthèse (Alcamo *et al.*, 2007), ce qui a pour conséquence une instabilité des rendements. La production céréalière en Australie ne cesse de fluctuer depuis 2003 en raison des épisodes de sécheresse. Récemment, les prix des denrées alimentaires ont augmenté menaçant la sécurité alimentaire. En 2010, les prix du blé ont ainsi subi une hausse de 50% suite à des vagues de chaleurs sans précédents en Russie (Teixera, 2011). En 2012, 64% du territoire des Etats-Unis a été frappé par une sécheresse historique, entraînant une flambée des prix du soja, maïs et blé. Les régions du globe sont différemment affectées par le réchauffement climatique. Même si certaines zones aux latitudes septentrionales semblent pouvoir bénéficier d'un impact positif, les effets attendus sont la plupart du temps néfastes dans les zones tropicales et subtropicales, souvent très vulnérables étant donnée leur dépendance à l'agriculture ainsi que leur forte croissance démographique. Il est probable que les surfaces affectées par la sécheresse augmentent. Plusieurs scénarios attestent d'une augmentation de 1 à 3°C sur la plupart des zones géographiques d'ici à 2050 (CABI, 2010), accompagnée de changements des régimes de précipitations, avec certaines régions victimes de sécheresses intenses. La sécurité alimentaire dépendra donc notamment de la capacité des cultures à tolérer des stress abiotiques de plus en plus fréquents et à maintenir leur rendement aussi stable que possible quelque soient les conditions.





Comparé à d'autres cultures d'été, le tournesol, cultivé principalement à bas intrants (produits phytosanitaires mais aussi eau et azote), est connu pour sa tolérance à la sécheresse grâce à un enracinement profond et à une capacité d'ajustement de ses surfaces foliaires sous contraintes hydriques. Cependant, des simulations ciblées sur la zone du bassin méditerranéen (Figure I.1) ont montré qu'il existait un risque important de pertes de rendement dû au réchauffement climatique. Ces pertes sont prédites en moyenne autour de 13%, et pourront aller jusqu'à 35% en prenant en compte les événements climatiques extrêmes, assimilés dans cette étude à des stress thermiques pendant l'anthèse (Moriondo *et al.*, 2011).

Disposer de variétés qui limitent leur perte de productivité lorsque les conditions hydriques et les températures deviennent contraignantes, constitue donc un réel enjeu.

## **I.2 Quel type de « contrainte hydrique » prendre en compte ?**

Il est important de distinguer la capacité d'une plante à survivre dans des conditions arides de celle permettant de maintenir un rendement aussi stable que possible dans des conditions de sécheresse faibles à moyennement intenses. Notre étude, comme la plupart des programmes de sélection, cible le deuxième type de contraintes. Dans ce contexte, la sécheresse survient dès que la demande en eau pour la transpiration excède la disponibilité en eau dans le sol. Les dommages physiologiques engendrés par la baisse de potentiel hydrique sont à l'origine du stress hydrique ressenti par la plante. Ce stress accélère, à des degrés divers, la fermeture des stomates, diminuant l'absorption de carbone, ce qui a pour conséquence de réduire le développement de biomasse et donc l'allocation des assimilats aux grains. La stérilité des organes reproducteurs peut aussi être observée dans certains cas. Le rendement, qui intègre tous ces processus, est donc directement impacté.

La température influençant également profondément le métabolisme des plantes, le stress thermique est aussi un facteur important à prendre en compte dans l'élaboration du rendement. Les stress thermiques dû à des températures élevées entraînent la réduction de la durée du cycle et donc le temps disponible pour intercepter le rayonnement et produire la biomasse. Ce stress n'est pas seulement lié à la température de l'air mais aussi à la température des feuilles, qui lorsqu'elle atteint des valeurs élevées, peut contraindre certains processus métaboliques. Quant au stress dû au froid, il provoque la déshydratation des tissus.

A travers les dommages qu'ils engendrent et les réponses communes, ces différents stress abiotiques sont fortement interconnectés.

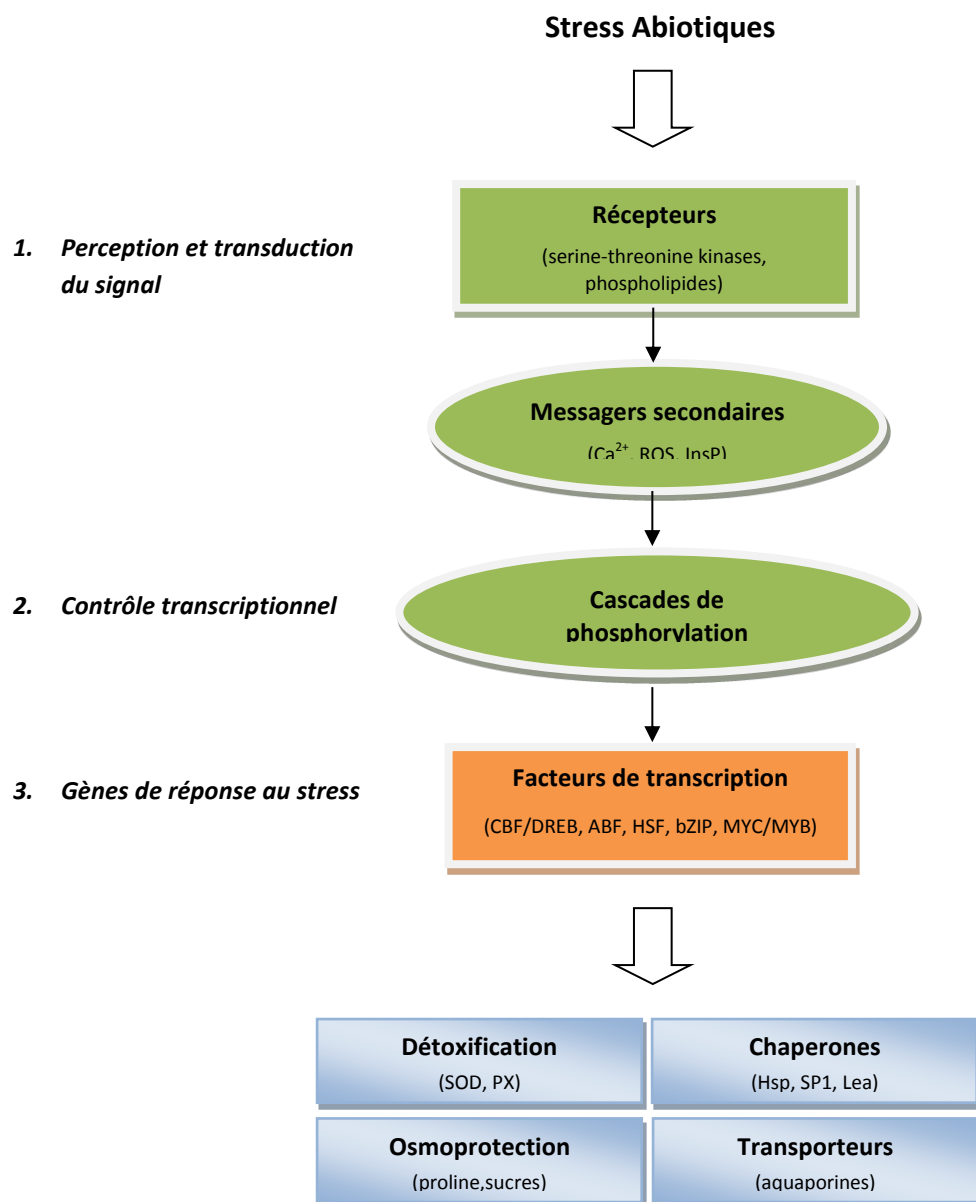


La suite de ce chapitre présente quelques résultats notables liés à la réponse des plantes cultivées au stress hydrique. Tous les aspects de cette réponse ne seront pas traités de manière exhaustive car nous avons choisi d'orienter cet état des lieux sur le thème de l'impact du stress sur la productivité, thème en lien avec le sujet de cette thèse.

### **I.3 Stratégies d'adaptation des plantes à la sécheresse**

Plusieurs stratégies ont été mises en place par les plantes pour compenser les effets négatifs dus au stress hydrique. Levitt (1980) les a classées en deux stratégies principales : l'évitement (« dehydration avoidance ») et la tolérance à la déshydratation (« dehydration tolerance »).

- L'évitement de la déshydratation vise à maintenir le statut hydrique de la plante malgré le stress hydrique présent. L'amélioration du prélèvement en eau, si les propriétés du sol le permettent (rôle des racines), de son stockage dans les tissus (conductance hydraulique) ainsi que la diminution des pertes et de la demande en eau (fermeture stomatique, réduction des surfaces foliaires et sénescence accélérée) en sont les principaux mécanismes. La fermeture des stomates diminue la transpiration, la captation de dioxyde de carbone et donc l'accumulation de biomasse, ce qui réduit inévitablement le rendement. Cette stratégie a donc un intérêt restreint en sélection pour la plupart des scénarios climatiques (Chenu *et al.*, 2009) excepté ceux où le stress est très intense (Yadav *et al.*, 2011). L'ajustement osmotique, (accumulation net de solutés tels que des ions, sucres et acides aminés) est également un des processus clés qui permet de maintenir le statut hydrique des cellules et donc la photosynthèse et la croissance mais à une vitesse plus lente.
- La tolérance à la déshydratation est définie comme la capacité de continuer à fonctionner avec un statut hydrique faible. La synthèse de certaines molécules telles que les antioxydants dirigés contre les ROS (« Reactive Oxygen Species »), les protéines « LEA » (« late embryogenesis abundant ») ou de choc thermique (« heat shock protein »), les carbohydrates, permet de protéger les fonctions physiologiques. La remobilisation des carbohydrates présents dans les tiges contribue également à compenser les effets négatifs du stress pendant la phase de remplissage du grain. Cette stratégie « productive » présente donc *a priori*, dans le cadre d'un stress modéré, un réel intérêt pour le sélectionneur.



**Figure I.2: Principaux mécanismes de la réponse aux stress abiotiques.** Ca<sup>2+</sup> : calcium, ROS : reactive oxygen species, InsP : inositol phosphate, CDPK : calcium-dependent proteines kinases, MAP : mitogen-activated, SOD : superoxide dismutase, PX: peroxidase, HSP : heat shock protein, LEA: late embryogenesis abundant (adapté de Wang et al., 2003).

En dehors de ces deux stratégies de nature physiologique, il faut noter par ailleurs l'existence d'un levier agronomique. En effet, le cycle de culture peut être adapté de façon à ce que les phases les plus sensibles au stress (autour de la floraison) ne coïncident pas avec les intensités de stress les plus fortes. Cette stratégie, communément appelé « échappement » (« drought escape »), entraîne le plus souvent une réduction de la durée du cycle (permettant d'abaisser la demande totale en eau), ce qui peut affecter le rendement potentiel.

En fonction de la période de survenue du stress, de son intensité et de sa durée, les plantes utilisent différents processus biochimiques, physiologiques et morphologiques liés aux stratégies précédentes. Les mécanismes moléculaires mis en jeu sont complexes et ne sont pas encore clairement élucidés.

#### **I.4 Mécanismes moléculaires impliqués dans la réponse au stress**

Nous n'aborderons cette question qui fait l'objet de très nombreuses recherches au niveau mondial - surtout sur plantes modèles et dans des conditions expérimentales très contrôlées - que de façon très synthétique. Schématiquement (Wang *et al.*, 2003), les gènes impliqués dans cette réponse peuvent être regroupés en trois types d'activité : la perception et la transduction du signal, le contrôle transcriptionnel et les mécanismes de réponse au stress à proprement parler tels que: la détoxification, l'osmoprotection, les aquaporines et les protéines chaperonnes protégeant la conformation structurelle des membranes ou fonctionnelle d'autres protéines.

La figure I.2 présente l'enchaînement de ces trois types d'activité liés à la réponse des plantes à différents stress abiotiques (hydriques, salins et thermiques). Les plantes ont en effet des voies de réponses communes à ces différents stress qui peuvent avoir pour même conséquence l'apparition de stress secondaires tels que les stress osmotiques et oxydatifs. La perception du signal, dans lequel la membrane plasmique joue un rôle prépondérant, est en général suivie de la production de messagers secondaires (ionisé phosphate ou ROS) qui modulent la concentration intracellulaire de calcium ( $Ca^{2+}$ ). Des protéines kinases dépendantes ou non du calcium (Boudsocq et Sheen, 2012) catalysent ensuite une cascade de réactions de phosphorylation avant de cibler des facteurs de transcription comme CBF/DREB (Yamahushi-Shinozaki et Shinozaki, 2006) dont certains régulent l'expression des gènes majeurs de réponses au stress. Le produit de ces gènes contribue à l'adaptation des plantes aux stress. La transduction du signal associé au stress hydrique est ou non sous la dépendance de l'acide abscissique (Huang *et al.*, 2012), hormone qui est impliquée dans le contrôle de



l'ouverture des stomates et dans la croissance. Plus généralement, c'est un véritable réseau de signalisation impliquant différentes hormones qui joue un rôle dans la réponse aux stress abiotiques (Kohli *et al.*, 2013). Afin d'améliorer la compréhension de ces réseaux complexes il est nécessaire d'utiliser différentes méthodes d'analyses.

Le moyen le plus direct pour vérifier le rôle des allèles d'intérêt de ces gènes dans la tolérance au stress hydrique reste de les transférer par rétrocroisement ou par génie génétique, puis de comparer les phénotypes de ces plantes quasi-isogéniques. Les facteurs de transcription constituent de bons candidats pour ce type d'approches. Parmi les exemples de réussite, on peut citer les plantes pour lesquelles les facteurs de transcription DREBs/CBFs ont été transférés par transgénèse tels que la tomate (Hsieh *et al.*, 2002), le riz (Ito *et al.*, 2006) et le blé (Pellegrineschi *et al.*, 2004). Ces plantes ont en effet augmenté leur tolérance au stress hydrique, tout comme *Arabidopsis* ou le tabac ayant intégré HDG11 qui code pour un facteur de transcription HD-STARTtype, (Yu *et al.*, 2008). L'analyse de ces plantes transgéniques est cependant souvent basée sur la capacité des plantes à survivre à un stress brutal sur une période très courte, ce qui est assez éloigné des objectifs visés en général par les programmes d'amélioration génétique des plantes de grande culture (Cattivelli, 2008). De plus, peu d'études ont eu lieu en plein champ, à part quelques exceptions comme par exemple pour le facteur de transcription SNAC1 chez le riz, qui a permis d'améliorer la tolérance au stress et le rendement au champ (Hu *et al.*, 2006). Le transfert de gènes est donc essentiel à la compréhension des mécanismes de réponse au stress mais se heurte à différentes limites, notamment au fait qu'il ne permet pas d'identifier tous les composants des voies métaboliques (Huang *et al.*, 2012).

Récemment, d'autres méthodes d'analyses, telles que la transcriptomique, ont été appliquées à l'échelle du génome entier, pour révéler d'autres gènes de réponse aux stress abiotiques. D'abord adoptée chez *Arabidopsis* (Kreps *et al.*, 2002 ; Matsui *et al.*, 2008), ces méthodes ont été transférées à d'autres espèces non modèles. Pour le tournesol, la plupart des connaissances sur les mécanismes moléculaires de réponse au stress hydrique proviennent des analyses du transcriptome. Voici quelques résultats obtenus :

- L'accumulation de transcrits en fonction du statut hydrique a été mesuré chez 2 lignées : une tolérante à la sécheresse grâce au maintien de la turgescence et l'autre sensible (Cellier *et al.*, 1998). Pour un potentiel hydrique bas, le taux de déhydrine transcrits était de 5 à 9 fois plus élevée chez la lignée résistante pour les gènes





- HaDhn1 et HaDhn2 respectivement, la déhydrine jouant un rôle important de la stabilisation des protéines et la tolérance au stress hydrique.
- Le gène Hahb-4, appartenant à la sous famille des protéines homeodomaine-leucine zipper, présente que chez le tournesol, est régulé au niveau transcriptionnel par la disponibilité en eau et l'acide abscissique (Dezar *et al.*, 2005). Son implication dans la réponse développementale de la plante à la dessiccation a été montrée, ainsi que dans le métabolisme de l'éthylène induisant un retard de la sénescence (Manavella *et al.*, 2006).
- Un microarray de 800 gènes (Roche *et al.*, 2007) recouvrant les principales voies métaboliques a permis d'identifier de nombreux gènes différentiellement exprimés dans les feuilles et embryons de génotypes tolérants ou sensibles exposés à des conditions de stress hydrique au champ. Les génotypes tolérants ont sur-exprimé des gènes impliqués dans la détoxification des cellules (aldehyde deshydrogenase, aminotransferases and tocopherol enzymes), tandis qu'ils ont sous-exprimé des gènes impliqués dans la division cellulaire, rappelant la stratégie d'évitement (croissance ralentie).
- Poormohammad Kiani *et al.*, (2007) ont sélectionné 4 RILs présentant des réponses au stress hydrique contrastées et leurs parents afin de déterminer l'expression de 4 gènes de réponse au stress : aquaporine, déhydrine, leafy cotyledon1- like protein et fructose-1,6 bisphosphatase. Ils ont pu montrer qu'il existait des corrélations entre les réponses physiologiques, les allèles aux QTL et l'expression des gènes. Ainsi le niveau d'expression des gènes d'aquaporine (sous-regulés lors du stress) était associé avec la teneur en eau (RWC) ainsi qu'avec les régions génomiques portant les allèles défavorables aux QTL.
- Une approche intégrant à la fois transcriptomique, physiologie et génétique permet d'améliorer la compréhension des mécanismes de tolérance au stress hydrique. Rengel *et al.*, (2012) ont calculé les corrélations entre l'expression des gènes présents sur une puce Affymetrix de 32 423 ensembles de sondes associés à des EST et la variabilité génétique de 8 génotypes pour 9 caractères morpho-physiologiques impliqués dans la réponse au stress. Les auteurs ont pu identifier des gènes et des processus physiologiques qui permettaient d'expliquer les différences de réponse au stress. De



plus, 84 gènes candidats ont pu être identifiés à la fois en serre et au champ sur des génotypes en combinaison hybride.

De plus en plus, les nouvelles technologies de séquençage (NGS) comme le RNAseq par exemple promettent - sous réserve que les méthodes d'analyses statistiques de ces données acquièrent une certaine maturité - d'éviter certaines limitations des microarray (telles que la nécessité de définir un « probeset » au préalable) et permettent donc de générer des données encore plus exhaustives. Dans tous les cas, même si la transcriptomique est un outil essentiel, notamment pour déterminer des réseaux de gènes, le manque de corrélation entre le nombre de transcrits et l'effet des gènes nécessite des validations (Kantar *et al.*, 2011).

Giordani *et al.*, (2010) présentent une approche différente pour identifier des gènes candidats probants pour la tolérance à la sécheresse. A partir de 8 gènes connus pour leur implication dans la réponse au stress, présents sous la forme d'une unique copie chez le tournesol, les auteurs ont étudié la variabilité de séquence de ces gènes sur 8 lignées d'origines diverses et montrant des phénotypes différents en relation avec le stress. Les indices de diversité ont mis en évidence des contraintes d'évolutions différentes selon les gènes, ce qui présage de leur rôle dans la réponse au stress hydrique.

De nombreuses connaissances ont donc été acquises sur les mécanismes moléculaires de réponse aux stress abiotiques chez les plantes. La complexité de cette réponse en fait un sujet de recherche intense pour lequel de nouvelles méthodes d'analyses se développent (par exemple les modèles graphiques de régulation génique/Gene Regulatory Network). Pourtant, ces connaissances ont été relativement peu exploitées jusqu'à ce jour en amélioration des plantes cultivées pour leur tolérance à la sécheresse ; par contre différents caractères constitutifs ou en relation avec la réponse au stress ont été pris en compte.

### **I.5 Résumé des caractères contrôlant la réponse des plantes au stress hydrique**

Le rendement figure le plus souvent parmi les objectifs principaux d'un programme de sélection. Il est le résultat du cycle de culture et traduit donc la capacité des plantes à résister aux stress abiotiques. D'un point de vue physiologique, la compréhension des déterminismes du rendement a fait l'objet de nombreuses études, notamment chez les céréales, qui ont abouti à la construction de différents modèles et ont largement permis d'établir un cadre théorique pour les sélectionneurs et généticiens (Salekdeh, 2009), leur permettant ainsi d'identifier des caractères cibles de la sélection.



L'équation de Passioura (1977) est globalement définie comme la quantité de matière sèche produite par unité d'eau perdue par évapotranspiration.

$$\mathbf{Yield = WU * WUE * HI}$$

Dans cette équation, le rendement est ainsi exprimé comme le produit de la quantité d'eau utilisée, (WU : water use), de la conversion de cette eau en biomasse sèche (WUE : water use efficiency) et de l'indice de récolte (HI : harvest index), qui rend compte de l'allocation de cette biomasse au produit de la récolte.

Le terme "WUE" est complexe et revête plusieurs définitions selon l'échelle utilisée. Initialement ratio du rendement sur l'irrigation (« more crop per drop » ; Kijne *et al.*, 2003), les physiologistes l'ont défini à l'échelle de la feuille comme le ratio de la photosynthèse sur la transpiration. Chaque terme de l'équation permet d'identifier plusieurs caractères à cibler pour améliorer le rendement sous conditions hydriques.

Les caractères cibles peuvent être classés en deux types : les caractères constitutifs, contrôlés par des gènes qui ne nécessitent pas la présence de stress pour leur expression, et les caractères d'adaptation au stress, mobilisés en réponse au stress hydrique (Blum, 1997).

#### **Parmi les caractères constitutifs :**

- **La surface foliaire totale** module la demande évaporative à l'échelle de la plante entière et donc la transpiration, ce qui la relie directement au terme « WU » de l'équation de Passioura. Ce caractère peut être mesuré à l'aide du LAI (Leaf Area Index), qui correspond à la surface foliaire active par unité de surface de sol. L'évapotranspiration augmente avec le LAI jusqu'à atteindre un plateau. Le LAI augmente lors de la croissance de la plante puis diminue avec la sénescence. Lorsque le stress hydrique est précoce (avant l'anthèse), on peut observer le développement de feuilles plus petites (LAI peu élevé) afin de limiter la demande évaporative. Connor *et al.*, (1985) ont obtenu les mêmes conclusions chez le tournesol. Pereyra-Irujo *et al.*, (2008) ont étudié 18 lignées de tournesol représentant la variabilité du matériel cultivé. La croissance foliaire a été modélisée à plusieurs niveaux (cellule, feuille et plante) en réponse à un stress appliqué sur le long terme. Leurs résultats suggèrent que les différences génétiques de taux de croissance foliaire sous contraintes hydriques pourraient provenir des propriétés des membranes cellulaires tandis que l'allongement de la durée de



croissance proviendrait du prolongement de la division cellulaire. Ces différents mécanismes physiologiques pourraient ainsi être combinés pour améliorer le rendement.

- **L'architecture racinaire** et en particulier la profondeur des racines, qui permet d'explorer davantage le sol, est un caractère essentiel pour le prélèvement en eau (lien avec le terme « WU ») et donc la transpiration. Cependant, l'intérêt de cibler ce caractère doit être envisagé au vu des propriétés du sol et du coût métabolique que génère le développement d'un système racinaire plus dense. Bolaños et Edmeades (1993) ont par exemple montré que la biomasse racinaire pouvait être corrélée négativement au rendement à partir d'une population de maïs tropical.

- **Le phénotype « stay green »** est caractérisé par le retard de la sénescence, processus normal de vieillissement de la plante et de remobilisation des ressources vers les organes reproducteurs. Les génotypes ayant un « stay green » important maintiennent plus longtemps la chlorophylle dans les tissus foliaires et donc l'activité photosynthétique et la transpiration. Le rendement peut être ainsi, au moins théoriquement, préservé sous contrainte hydrique. Ce caractère est donc à la fois relié au terme WU pour la transpiration, au terme WUE pour l'activité photosynthétique et au terme HI pour la remobilisation. Chez le sorgho, il a été montré que le caractère « stay green » procure un avantage lors d'un stress post-anthèse (Borrell, 2000). Le « stay green » fonctionnel (retard de la perte de capacité photosynthétique) a été également démontré chez le tournesol et pourrait contribuer à maintenir le rendement stable en conditions hydriques limitantes (Gimenes et Fereres *et al.*, 1986 ; Iosada *et al.*, 1992 ; De la Vega *et al.*, 2011). Cependant, il existe un « trade-off » pour ce caractère : Hammer *et al.*, (2005) ont simulé l'effet du caractère « stay green » sur le rendement dans 547 environnements et ont montré que cet effet était positif lorsque le stress survient à partir du milieu de la saison mais négatif quand il survient de manière trop sévère en fin de saison. En effet, la présence d'un appareil végétatif actif entraîne une demande évaporative plus forte.

- **La synchronisation du cycle de la plante avec le stress hydrique** a été largement utilisée en sélection. Plusieurs succès ont pu être obtenus (Ludlow *et al.*, 1990) à partir du décalage de la date de floraison, dû en partie à l'héritabilité forte de ce caractère et sa facilité de mesure. Chez le tournesol notamment, l'avancée des dates de semis a permis d'améliorer les rendements dans les environnements agronomiques qui s'y prêtaient (De la Vega, 2002 ; Soriano *et al.*, 2004). Ce type d'approche peut cependant présenter un inconvénient dans le





cas d'environnements moins stressés où une durée courte de cycle entraîne des rendements plus faibles car il y a moins de temps disponible pour intercepter la lumière et produire de la biomasse. C'est donc un caractère qui doit être adapté au stress subi dans chaque environnement.

### **Parmi les caractères adaptatifs :**

- **Le caractère ASI** (anthesis silking interval), intervalle entre les floraisons mâles et femelles chez le maïs, est influencé par le stress hydrique qui entraîne un retard de la floraison femelle. L'émission de pollen ne coïncidant pas avec la réceptivité des soies, les conséquences peuvent être néfastes pour le rendement, ceci à travers l'impact sur l'indice de récolte (terme HI). La sélection pour minimiser cet intervalle, a permis d'améliorer le rendement en condition de stress hydrique (Ribaut et al, 2004). L'ABA aurait un rôle dans la tendance de certains géotypes à limiter le développement des organes reproducteurs lors d'un stress hydrique (Setter, 2012).

- **La discrimination isotopique naturelle du carbone (delta 13C)** est un indicateur de l'efficacité d'utilisation de l'eau (WUE) dans les feuilles (Condon *et al.*, 2002). Il existe plusieurs isotopes du carbone présents naturellement, dont l'un  $^{12}\text{C}$  est plus abondant que l'autre  $^{13}\text{C}$ . Les plantes, qui ont tendance à discriminer le carbone  $^{13}\text{C}$  en faveur du  $^{12}\text{C}$  plus léger, diffusant plus vite à travers les pores, et pour lequel l'enzyme Rubisco (fixation du carbone) a une meilleure affinité, présentent un rapport isotopique  $^{13}\text{C}/^{12}\text{C}$  plus faible que l'air. Ce caractère présente cependant des corrélations très variables avec le rendement : de positives lorsque beaucoup d'eau est disponible à négatives en conditions de stress intense. Ce caractère a été utilisé pour des conditions de stress sévères, par exemple en Australie où il a permis de créer des variétés de blés plus tolérantes aux conditions de sécheresses locales (Condon *et al.*, 2004) mais peut se révéler être un caractère limitant pour le rendement lorsque les conditions hydriques sont optimales (Tardieu *et al.*, 2011).

- **L'ajustement osmotique (AO)**, a également prouvé son utilité comme critère de sélection pour l'amélioration du rendement. Il permet de maintenir la turgescence des cellules et donc de contrôler la conductance stomatique et d'améliorer la capacité d'extraction dans les couches profondes du sol par les racines (terme WU). Sur le tournesol, Chimenti *et al.*, (2002) ont montré que l'AO permettait d'améliorer le rendement de 30% en conditions de stress pré-floraison. La demande énergétique requise pour l'AO peut cependant entraîner un



« trade-off », c'est-à-dire un compromis en terme de rendement lorsque les conditions de stress deviennent intenses (Kramer and Boyer, 1995).

- **La teneur relative en eau (RWC : relative water content)** mesure le statut hydrique des tissus car il correspond au volume d'eau contenu dans les feuilles relativement au volume maximum lorsque les tissus sont totalement turgescents. Il traduit donc l'impact de l'ajustement osmotique lors du stress hydrique et donc la capacité de la plante à éviter la déshydratation. Ce critère a été utilisé par H.Serieys à l'INRA de Montpellier pour conduire une sélection divergente au sein d'une population issue d'un croisement entre du tournesol cultivé et un écotype de l'espèce *H. argophyllus*. Une lignée issue de cette approche (AA.7.2.4) s'est révélée plus apte à réaliser une productivité correcte en condition de stress hydrique que d'autres matériels génétiques en comparaison.

Les quelques caractères détaillés ci-dessus présentent donc différents effets sur le rendement selon la nature du stress hydrique impliqué (intensité, occurrence au cours du cycle et durée). Liés principalement à la stratégie d'évitement (ex: fermeture des stomates), les caractères de réponse au stress présentent un désavantage lorsque les conditions hydriques sont favorables. Au contraire, les caractères constitutifs ne pénalisent pas le rendement en conditions favorables. La plupart des progrès réalisés sur le rendement en conditions hydriques limitantes provient des caractères constitutifs mesurés dans des conditions optimales (Chenu et al, 2009). Ainsi en maïs, le taux de progression du rendement dans la région du Corn Belt des Etats-Unis a été similaire en condition non stressée qu'en condition stressée (Tardieu *et al.*, 2011). Pour qu'un caractère soit une bonne cible pour l'amélioration génétique du rendement, il faut donc qu'il permette de stabiliser ce rendement en conditions de disponibilité en eau très diverses sans le pénaliser en conditions favorables, qu'il soit bien héritable et corrélé au rendement, qu'il présente suffisamment de variabilité génétique et qu'il soit intégratif afin de traduire une réponse à long terme sur un niveau d'organisation plus large (exemple : canopée). Enfin, il doit être également facilement mesurable. Actuellement, il existe encore beaucoup de difficultés liées à l'acquisition de certaines variables à grande échelle (telles que par exemple l'architecture racinaire). Les technologies à haut débit se développent de plus en plus pour le phénotypage (ex : la réflectance spectrale, Montes *et al.*, 2007) et promettent d'améliorer la précision des données et le suivi des caractères de manière dynamique.



En résumé, la stratégie la plus avantageuse pour l'objectif de sélection, est de maintenir la transpiration et la capacité de la plante à mieux utiliser l'eau disponible dans le sol.

### **I.6 Quels sont les progrès obtenus pour la tolérance à la sécheresse en utilisant la sélection assistée par marqueurs?**

Chez le maïs, le rendement en conditions de sécheresse a pu être amélioré comme le montre l'étude de Barker *et al.*, (2005) : 18 hybrides inscrits entre 1953 et 2001 ont été testés en conditions stressantes et le gain génétique a été estimé de 91 à 124 Kg/ha/an pour un stress autour et après floraison respectivement. La réduction de l'intervalle ASI (anthesis-silking interval) y a sans doute contribué fortement. Plusieurs QTL ont été identifiés pour ce caractère (dont un expliquant 38% de la variabilité phénotypique) et l'introgession de l'allèle favorable dans des variétés tropicales de maïs a permis d'améliorer le rendement lorsque le stress hydrique survient pendant la floraison sans le pénaliser en conditions non limitantes (Ribaut and Ragot, 2007).

Chez le riz, de nombreux QTL ont été détectés pour des caractères d'ajustement osmotique, de WUE, de phénologie (revue dans Vinod, 2006). Quatre QTL dont les allèles favorables conférant un enracinement plus profond ont été transférés de Azucena, une variété de type « japonica » pluviale à IR64, une variété de type « indica » (Courtois *et al.*, 2003). Les lignées introgressées à partir des QTL cartographiés finement développent des racines plus longues en conditions de stress hydrique (Steele *et al.*, 2006) et l'un de ces QTL améliorent même la capacité de pénétration des racines (Clarke *et al.*, 2008).

Chez le sorgho, plante réputée assez tolérante au déficit hydrique notamment autour de la floraison, quatre QTL majeurs ont été identifiés pour le caractère « stay green » (Harris et al, 2007). Les allèles favorables ont contribué à un retard de la sénescence en condition de stress post-anthèse.

Même si le blé sert de référence dans la compréhension des mécanismes physiologiques, il existe encore peu d'exemples d'amélioration de la tolérance au stress hydrique en sélection assistée par marqueurs, de part notamment la complexité du génome de cette espèce. Un exemple de succès cependant provient de l'utilisation de l'ajustement osmotique pour améliorer le rendement chez le blé. Ce caractère, très héritable (Moinudin *et al.*, 2005),



contrôlé par un gène majeur et quelques gènes mineurs (Morgan, 1991), a permis d'obtenir une variété tolérante au stress hydrique (Munns and Richards, 2007).

Chez le tournesol, il existe quelques exemples de détections de QTL sous contraintes hydrique au champ et en serre. Ebrahimi *et al.*, (2008) ont étudié l'effet du stress hydrique sur la teneur en huile ainsi que d'autres caractères de qualité des graines au champ et en serre, à partir d'une population de RILs (PAC2x RHA266). A part une région sur le LG16, la plupart des QTL détectés pour la teneur en huile était spécifique à une condition hydrique. De la même façon les QTL détectés pour les caractères liés à la photosynthèse et au statut hydrique (RWC, ajustement osmotique, potentiel hydrique) (Poormohammad Kiani *et al.*, 2007) diffèrent selon le traitement hydrique. Les auteurs ont montré des colocalisations entre les paramètres de photosynthèse et de statut hydrique. En utilisant la même population de RIL, Poormohammad Kiani *et al.*, (2009) ont détecté des QTL pour différents caractères agronomiques communs à quatre conditions (champ irrigué ou non/serre stressé ou non) ainsi que des colocalisations avec des QTL physiologiques de l'étude de Poormohammad Kiani *et al.* (2007). Les conditions climatiques n'avaient pas permis d'imposer un stress suffisant au champ. Ces résultats mettent aussi en évidence l'existence de QTL constitutifs (communs à plusieurs environnements) par rapport aux QTL adaptatifs (spécifiques à un environnement). L'importance de l'ajustement osmotique pour maintenir le rendement sous contraintes hydrique a été démontrée chez le tournesol (Chimenti *et al.*, 2002). Poormohammad Kiani *et al.*, (2007) ont localisé 8 QTL pour l'ajustement osmotique à partir d'une population de 129 RILs. Quatre d'entre eux colocalisent avec d'autres variables de statut hydrique. Un QTL majeur a été identifié sur le LG5, expliquant 29% de la variation phénotypique.

Il existe donc une variabilité génétique importante chez le tournesol pour les caractères utiles à l'amélioration du rendement sous contraintes hydriques. Cependant, il y a encore peu d'études visant à détecter les QTL impliqués dans ce trait, en dehors de quelques études en conditions contrôlées (serre) sur des effectifs faibles. Les conditions expérimentales sont donc assez éloignées du contexte « hybride de plein champ ».

Les exemples d'utilisation des marqueurs en sélection pour le rendement chez le tournesol sont quasiment inexistant, hormis celui de Eathington *et al.*, (2007). Les auteurs ont démontré l'efficacité de la sélection récurrente assistée par marqueur (SRAM) appliquée par l'entreprise de sélection Monsanto dans une population européenne de tournesol. Le principe de cette méthode est d'accroître la fréquence des allèles favorables en sélectionnant et croisant les plantes ayant les meilleures combinaisons alléliques aux QTL. Dans les résultats de cette entreprise, les lignées issues du schéma SRAM ont un potentiel de 10 Kg/ha de rendement

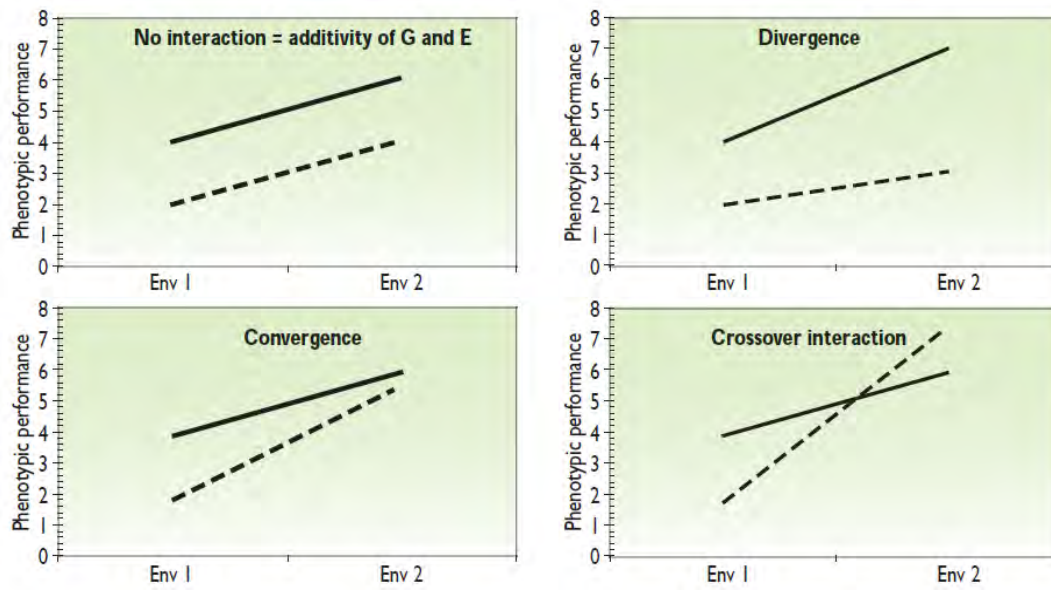




grains en plus que celles issues du programme conventionnel (sans l'utilisation des marqueurs moléculaires).

En résumé, beaucoup de QTL ont été identifiés chez la plupart des plantes pour de nombreux caractères physiologiques et morphologiques, constitutifs ou de réponses au stress (une base de données est d'ailleurs accessible à l'adresse: <http://www.plantstress.com/biotech/index.asp?Flag=1>) mais les cas de variétés résistantes obtenues par sélection assistée par marqueurs sont très rares. En dehors des questions liées à la confidentialité, une des raisons à cela peut être que les effets alléliques associés aux caractères complexes varient d'une population à l'autre (Jonas *et al.*, 2013). Les allèles sont en général détectés en conditions contrôlées, dans des populations expérimentales, non utilisées en sélection et les effets sont souvent très faibles. Ces contraintes ont amené d'autres méthodologies à se développer, telles que la sélection génomique (Meuwissen *et al.*, 2001), dont le principe est d'estimer la valeur d'un individu en utilisant un grand nombre de marqueurs moléculaires simultanément, et non seulement quelques marqueurs associés aux QTL détectés comme dans l'approche SRAM. La première étape consiste à établir une formule de prédiction des valeurs des individus à partir d'une population de calibration pour laquelle des données génotypiques et phénotypiques sont disponibles. Cette approche permet d'intégrer les effets à tous les marqueurs et donc de capturer une part importante de la variance génétique. Ensuite, des individus, de préférence reliés génétiquement à la population de calibration, sont génotypés mais pas phénotypés, leur valeur (GEBV : Genomic Estimated Breeding Values) étant calculée à partir de leur profil moléculaire. D'abord initiée chez l'animal (Hayes *et al.*, 2009), l'utilisation de la sélection génomique s'est développée chez les plantes car elle offre la possibilité d'accroître le gain génétique, notamment pour les caractères complexes.

Même si la sélection génomique offre d'ores et déjà des preuves de son efficacité (Cabrerat-Bosquet *et al.*, 2012), un des aspects liés aux difficultés de la sélection assistée par marqueurs n'est pas résolu pour l'instant. Il s'agit de l'influence de l'environnement sur l'effet du marqueur détecté. Les allèles peuvent être plus ou moins favorables selon le scénario de stress rencontré (Collins *et al.*, 2008). La question essentielle est de savoir si un allèle donné confère un effet positif sur le rendement dans le maximum de scénarios climatiques rencontrés dans les environnements ciblés (TPE : Target population of environments) (Tardieu *et al.*, 2011). L'interaction génotype - environnement (GE) contraint la tenue de cet objectif. On parle d'interaction GE lorsque les performances relatives entre génotypes varient selon les environnements. L'interaction GE se manifeste par des corrélations génétiques faibles entre



**Figure I.3: Interaction génotype-environnement à travers les changements de performances entre deux environnements (Env 1 et Env 2) (Monneveux & Ribaut, 2011).**

les environnements et/ou des variances génétiques hétérogènes selon ces environnements. Les observations phénotypiques de géotypes sur un environnement ne permettent donc pas de prédire les performances de ces géotypes sur un autre environnement. La figure I.3 représente la performance de 2 géotypes en fonction des environnements (normes de réaction) dans plusieurs situations où l'interaction GE est présente. Une des situations les plus critiques est le changement de classement (crossover interaction).

Il est donc nécessaire de prendre en compte l'interaction GE, d'autant plus marquée dans un contexte de variabilité des conditions hydriques, afin de prédire avec plus de précision la performance des géotypes. Pour cela, de nombreuses stratégies ont été proposées dans la littérature.

### **I.7 Stratégies d'analyses de l'interaction GE**

Il existe trois approches principales pour gérer l'interaction GE :

- « l'ignorer » : cela ne signifie pas que l'on considère l'interaction GE absente, mais plutôt, que l'on cherche à tester les géotypes sur le plus grand nombre d'environnements de manière à avoir la performance moyenne la plus forte. Les analyses de variance permettant d'estimer la variance de l'environnement et de l'interaction GE peuvent être utilisées dans cette approche pour évaluer la réponse à la sélection.
- « la réduire » : il s'agit de diviser la population d'environnements de culture cible TPE en petit groupes d'environnements aux caractéristiques similaires (par exemple, environnementales) et dans lesquels l'interaction GE est diminuée. Des analyses multivariées, de type ACP (analyses en composantes principales), sont en général un moyen d'y parvenir mais ces classifications sont parfois difficiles à interpréter du point de vue agronomique.
- « l'exploiter » : en plus d'exploiter la performance moyenne sur un ensemble d'environnements (comme dans la 1<sup>ère</sup> approche), cette méthode vise aussi à identifier les meilleures combinaisons géotype-environnements. Cela peut se traduire par la mise en œuvre de modèles multiplicatifs (type AMMI) ou des analyses de stabilité qui permettent d'appréhender la performance des géotypes comme une fonction du niveau de productivité (qualité) des environnements ou d'autres facteurs agronomiques limitant (ex : degré de pression d'un agent pathogène ou d'un stress abiotique).



Après un bref rappel de l'analyse de variance classique dans le cadre de la première approche, nous nous centrerons sur la 3<sup>ème</sup> approche visant à exploiter l'interaction GE, et en particulier sur les analyses de stabilité.

Les premières méthodes statistiques d'analyse de l'interaction GE sont basées sur l'analyse de variance à partir du modèle linéaire classique suivant :

$$Y_{ij} = \mu + G_i + E_j + GE_{ij} + \varepsilon_{ij}$$

où  $Y_{ij}$  est le phénotype de l'individu  $i$  dans le lieu  $j$ ,  $\mu$  est la moyenne générale,  $G_i$  est l'effet génétique de l'individu  $i$ ,  $E_j$  est l'effet de l'environnement  $j$ ,  $GE_{ij}$  est l'interaction entre le génotype  $i$  et l'environnement  $j$ ,  $\varepsilon_{ij}$  est l'erreur du modèle. Dans le cadre des modèles à effets fixes, ce modèle n'est pas estimable car sur-paramétré. Avec la généralisation des modèles mixtes, l'interaction GE et un des effets principaux (soit l'effet environnement, soit l'effet génotype) sont considérés comme aléatoires (Smith 2005). Ce modèle présente cependant une limite majeure. Les effets d'interaction aléatoires ne peuvent être prédits que si une information *a priori* permet de séparer la variabilité due à la résiduelle de celle due à l'interaction. Pour palier à cette limitation, la plupart des modèles statistiques proposés par la suite vise à modéliser le terme GE afin de réduire son nombre de niveaux.

Le modèle AMMI (additive main effects and multiplicative interaction) (Denis et Vincourt, 1983 ; Gauch, 1988) est un modèle dans lequel les effets additifs ( $G_i$  et  $E_j$ ) sont estimés par une analyse de variance tandis que l'interaction GE est modélisée de façon multiplicative et analysée à travers une ACP.

$$Y_{ij} = \mu + G_i + E_j + \sum_{k=1}^K b_{ik} Z_{jk} + \varepsilon_{ij}$$

Dans ce modèle, le terme GE est ici expliqué par  $K$  termes multiplicatifs, chacun étant le produit d'une sensibilité génotypique  $b_{ik}$  (score génotypique) et d'une caractérisation environnementale  $Z_{jk}$  (score environnemental). Ces scores sont les composantes principales respectives des génotypes et des environnements sur le  $k$  ième axe principal d'une ACP. Le modèle AMMI présente aussi l'intérêt, à travers une visualisation sous forme de « biplot », d'identifier les génotypes particulièrement adaptés à un environnement spécifique. Cependant, l'interprétation biologique de ces modèles n'est pas toujours évidente et ils ne permettent pas de prendre en compte des variables environnementales (climat, sol...).



Les analyses de stabilité regroupent toutes les méthodes visant à examiner la réponse d'un génotype sur un ensemble d'environnements, et particulièrement sa capacité à maintenir son rendement stable. Parmi les méthodes utilisées, la « régression factorielle » linéaire a l'avantage d'inclure des covariables environnementales. Finlay et Wilkinson (1963) ont, les premiers, décrit l'interaction GE à partir d'une régression de la performance sur le niveau moyen de chaque environnement:

$$Y_{ij} = \mu + G_i + E_j + b_i E_j + \varepsilon_{ij}$$

où  $b_i E_j$  mesure donc le changement de performance d'un génotype par unité de changement de « niveau » environnemental, c'est donc un indicateur de stabilité sur l'ensemble du réseau expérimental. C'est pourquoi celui-ci doit être suffisamment large et avoir une distribution uniforme des environnements, en termes de niveau, pour pouvoir correctement estimer la pente de la régression. Une mesure de « qualité » environnementale est donc nécessaire pour distinguer et classer les environnements, ce dont dépendra l'adéquation du modèle. Les sélectionneurs ont longtemps caractérisé les réseaux en utilisant la performance moyenne de tous les génotypes présents sur l'environnement ou de variétés témoins. Grâce à la diminution des coûts d'acquisition des données, de plus en plus de variables environnementales (bilan hydrique, sol...) sont maintenant utilisées pour décrire les environnements. Hodson et White (2007) ont ainsi utilisé des données relatives au climat et au sol pour caractériser leur réseau ; Hernandez-Segundo *et al.*, (2009) ont utilisé des caractères physiologiques et Chapman (2000) des index de stress issus de modèle de culture. Chenu *et al.*, (2011) ont regroupés les environnements d'un réseau d'essai de blé en Australie selon les scénarios de stress rencontrés en utilisant le modèle de culture APSIM (Agricultural Production System Simulator). L'intégration de cet effet « scénario de stress » dans les modèles statistiques a permis de réduire l'effet de l'interaction GE par rapport à l'effet génétique dans l'explication du phénotype observé. L'intérêt des modèles de culture pour caractériser les environnements a été démontré (Muchow *et al.*, 1996), de même que pour modéliser plus précisément le phénotype, mais peu d'études vont jusqu'à utiliser ces modèles pour identifier des régions génomiques.

Pour identifier les bases génétiques de l'interaction GE, l'utilisation de modèles mixtes intégrant l'effet du QTL et celui de l'interaction QTL x Environnement, a été plus largement répandu (Boer *et al.*, 2007). Cependant, d'autres stratégies ont été mises en place, s'appuyant davantage sur la physiologie, telle que la « cartographie fonctionnelle » (Van Eeuwijk *et al.*, 2010), le caractère à cartographier n'étant pas le caractère *per se* mais plutôt les paramètres





d'une courbe de réponse de ce caractère en fonction de différentes conditions environnementales. Cette méthode a été appliquée à des caractères relevant le plus souvent de processus biologiques spécifiques (Reymond *et al.*, 2003). A l'échelle des caractères complexes, Zheng *et al.*, (2010) ont identifié des QTL de sensibilité des géotypes à partir de la pente de la régression de composantes du rendement sur des covariables environnementales liés au stress hydriques et azotés (sur des variétés témoins) mais non issues d'un modèle de culture. Dans le cadre de cette thèse, nous avons choisi d'utiliser un modèle de culture, spécialement construit pour modéliser la performance de géotypes de tournesol en lien avec leur environnement (SUNFLO, décrit dans le chapitre II) afin de caractériser un réseau d'essais et de décrire l'interaction géotype - environnement.

En résumé, dans un contexte de disponibilité en eau incertain pour les terres agricoles, la stabilité du rendement est plus que jamais un caractère essentiel sur lequel les efforts de sélection doivent porter. La tolérance au stress hydrique ne doit pas se faire au détriment du rendement. Il est donc nécessaire de capitaliser sur les stratégies de fonctionnement des plantes les plus productives c'est-à-dire le prélèvement en eau et la transpiration. Malgré l'amélioration de la compréhension des mécanismes morphologiques, physiologiques et moléculaires à la base de la tolérance au stress hydrique, les progrès génétiques notamment grâce à la sélection assistée par marqueurs se font attendre. Hormis la complexité du phénotype, c'est aussi l'interaction GE qui constitue un réel frein à ce progrès. Celle-ci doit être donc traitée conjointement avec l'étude du rendement.

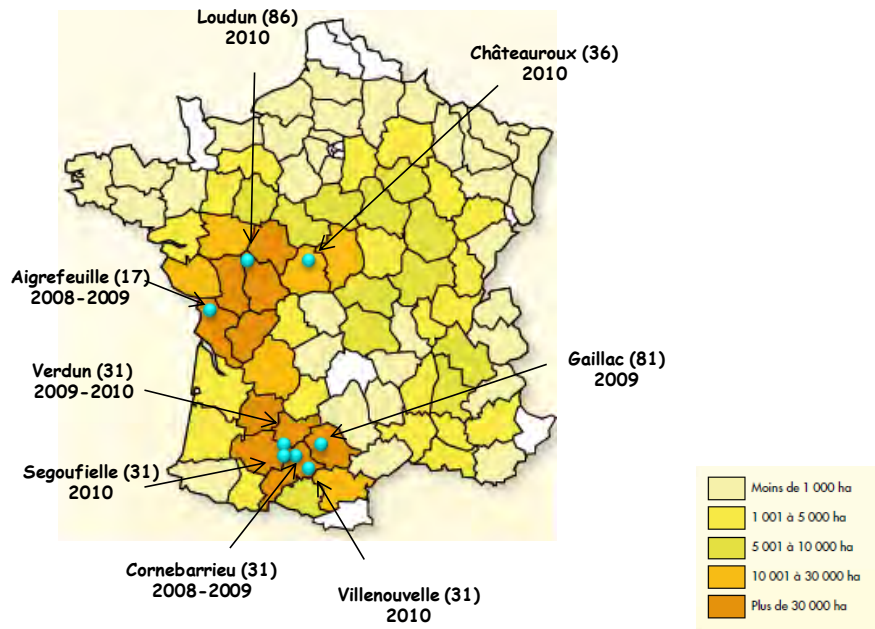


Fig II.1 : Répartition des lieux dans les zones de production de tournesol (sources : Onidol, 2010).

## Chapitre II Analyse des données phénotypiques

### II.1 Introduction

Dans le cadre de cette thèse, de nombreuses données phénotypiques ont été acquises. Grâce à la diversité des caractères mesurés et des environnements expérimentés, nous avons la possibilité de comprendre davantage le comportement des génotypes dans des situations de contraintes hydriques diverses. L'objectif de ce chapitre est, dans un premier temps, d'analyser les données phénotypiques par environnement afin d'obtenir la meilleure estimation possible des valeurs génétiques. Puis, après l'étude des corrélations entre caractères, nous nous attacherons à caractériser le stress hydrique subit dans chaque environnement grâce à un modèle de culture. Enfin la performance des génotypes sera reliée à cette caractérisation, afin d'obtenir de nouveaux critères fondés sur le comportement à l'échelle multilocale permettant d'illustrer la variabilité des réponses génétiques au stress hydrique. C'est sur ces critères dont ce chapitre aboutit à l'élaboration que sera plus loin conduite une partie de l'étude de génétique d'association.

### II.2 Analyse des données phénotypiques par environnement

#### II.2.1 Matériels et méthodes

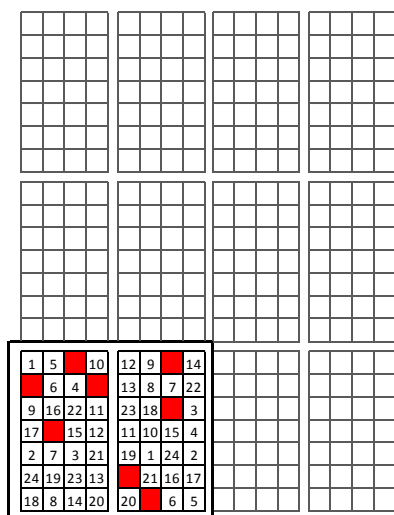
##### II.2.1.1 Matériel végétal

Le panel initial d'association rassemble 384 lignées publiques ou privées de type cultivé, appartenant à la collection des lignées de l'INRA ou impliquées dans des hybrides commercialisés (élites), dont certaines comprennent des introgressions d'*Helianthus* sauvages. La composition du panel et sa structure sont décrits dans le chapitre III.

Les 384 lignées du panel d'association ont été évaluées en plein champ de 2008 à 2010 sur un dispositif multilocal situé dans les zones de productions du tournesol en France (Figure II.1). Les lieux correspondant aux années 2008 et 2009 sont caractérisés par la présence de deux conditions de culture dans chaque lieu: une condition irriguée (apports d'eau appliqués autour de la floraison) et une condition sans irrigation. Certains lieux évalués en 2010 ont été irrigués si nécessaire. Chaque combinaison lieu-traitement (irrigué ou non) est appelée

Environnement	Lieu	Année	Condition	Testeur pour lignées B	Testeur pour lignées R	Nombre d'hybrides évalués
AI08_I	Aigrefeuille	2008	irrigué	83HR4gms	FS71501	247
AI08_NI	Aigrefeuille	2008	non irrigué	83HR4gms	FS71501	247
CO09_I	Cornebarrieu	2009	irrigué	83HR4gms	FS71501	349
CO09_NI	Cornebarrieu	2009	non irrigué	83HR4gms	FS71501	349
GA09_I	Gaillac	2009	irrigué	83HR4gms	FS71501	346
GA09_NI	Gaillac	2009	non irrigué	83HR4gms	FS71501	346
LO10	Loudun	2010	non irrigué	83HR4gms	FS71501	359
VE10	Verdun	2010	Irrigué	83HR4gms	FS71501	370
AI09_I	Aigrefeuille	2009	Irrigué	SOLR001M	AT0521	343
AI09_NI	Aigrefeuille	2009	non irrigué	SOLR001M	AT0521	343
VE09_I	Verdun	2009	Irrigué	SOLR001M	AT0521	336
VE09_NI	Verdun	2009	non irrigué	SOLR001M	AT0521	336
CA10	Castelnaudary	2010	non irrigué	SOLR001M	AT0521	370
CO08_I	Cornebarrieu	2008	Irrigué	SOLR001M	AT0521	299
CO08_NI	Cornebarrieu	2008	non irrigué	SOLR001M	AT0521	311
SE10	Segoufielle	2010	non irrigué	SOLR001M	AT0521	311
CHA10	Châteauroux	2010	Irrigué	SOLR001M	AT0521	370

**Table II.1 : Description des environnements**



**Figure II.2 : Dispositif expérimental : Les génotypes 1 à 24 du panel sont randomisés dans 2 blocs contenant chacun 4 témoins (en rouge). Les autres blocs sont organisés de la même façon avec les autres lignées du panel.**

« environnement ». Au total, le panel a été phénotypé dans 17 environnements dont les caractéristiques sont décrites dans la section II.3.3.

Une caractéristique importante du dispositif expérimental est que, de manière à s'affranchir des contraintes biotiques ou de caractères indésirables liés à la présence de matériel sauvage qui pourraient biaiser l'évaluation ou affecter la variabilité des caractères mesurés, le panel a été évalué en valeur hybride, grâce à l'utilisation de testeurs présentant une bonne aptitude générale à la combinaison. Dans un environnement donné, chaque lignée est croisée avec un testeur selon qu'elle est de type B (mainteneuse de la stérilité mâle) ou R (restauratrice de la fertilité) (Table II.1). Deux testeurs différents par lignée ont été utilisés selon les lieux. Ainsi les lignées B ont été croisées avec 83HR4gms dans 8 environnements et SOLR001M dans 9 environnements. 83HR4gms est issue de la conversion de 83HR4, une lignée restauratrice de fertilité, avec une stérilité mâle génique qui permet de l'utiliser comme femelle en croisement. SOLR001M est une lignée privée restauratrice de la stérilité mâle cytoplasmique classique (PET1) mais convertie en femelle sur le cytoplasme PEF1 (Crouzillat *et al.*, 1991), d'origine *Helianthus petiolaris sp. fallax*, dont elle maintient la stérilité mâle. Les hybrides issus du croisement d'une lignée de type B avec le testeur SOLR001M sont en général mâles stériles. Les hybrides entre le testeur (CmsPET1) FS71501 et les lignées R ont été testés dans 8 environnements tandis que ceux entre le testeur (Cms PET1) AT0521 et les lignées R l'ont été dans 9 environnements.

Plusieurs variétés, dont certaines connues des sélectionneurs et caractérisées pour les paramètres génétiques du modèle de culture SUNFLO (Casadebaig *et al.*, 2008) - ce qui permettra de les utiliser comme variétés indicatrices de l'environnement (Chapitre III) - ont été utilisées comme témoins dans les essais. Il s'agit de INEDI, issu du croisement XRQ\*PSC8 (croisement qui a également permis de générer une population de lignées recombinantes utilisée dans le cadre de cette thèse, cf. chapitre V), de MELODY, TEKNY et PEGASOL qui ont été des variétés largement cultivées.

#### II.2.1.2 Dispositif et caractères mesurés

Le dispositif d'expérimentation est organisé de la manière suivante (Figure II.2). Dans chaque environnement, le panel est divisé en blocs de 24 à 30 génotypes, répétés en deux sous-blocs randomisés et contenant chacun les 4 hybrides témoins. Ce dispositif appelé « augmented design » (Federer, 1956) est caractérisé par la répétition des témoins de nombreuses fois tandis que les autres génotypes ne sont pas répétés autant. Un ensemble de caractères phénotypiques a été mesuré sur chacun des génotypes (Table II.2). Parmi ces caractères, on

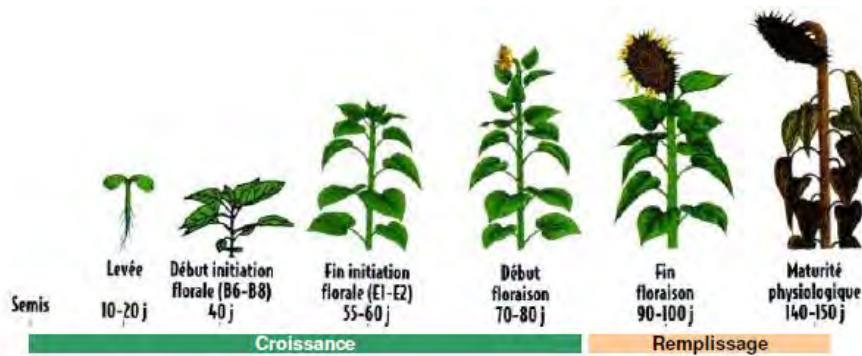


Figure II.3 : Stades phénologiques du tournesol (source : Cetiom)

Variable	Catégorie	Stade	Description
F1	Phénologie	-	Date de floraison (jours quantième)
M0	Phénologie	-	Début de remplissage du grain (jours quantième)
M3	Phénologie	-	Fin de remplissage du grain (jours quantième)
F1M0	Phénologie	-	Durée entre F1 et M0 (jours)
M0M3	Productivité	-	Durée entre M0 et M3 (jours)
H2O	Productivité	récolte	Teneur en eau des grains (%)-précocité récolte
Hauteur	Foliaire	F1	Hauteur totale des plantes (cm)
LAI_F1	Foliaire	F1	Leaf Area Index : index de surface foliaire
LAI_F20	Foliaire	F1+20 jours	Leaf Area Index: index de surface foliaire
LAI_F40	Foliaire	F1+40 jours	Leaf Area Index: index de surface foliaire
IFF1	Foliaire	F1+20 jours	Indice foliaire : échelle de 0 à 9, avec 9 =90% de plante verte
IFF20	Foliaire	F1+20 jours	Indice foliaire : échelle de 0 à 9, avec 9 =90% de plante verte
IFF30	Foliaire	F1+30 jours	Indice foliaire : échelle de 0 à 9, avec 9 =90% de plante verte
IFF40	Foliaire	F1+40 jours	Indice foliaire : échelle de 0 à 9, avec 9 =90% de plante verte
Flétrissement	Foliaire	Entre F1 et M3	0= pas flétri 1= 1/3 flétri 2= 2/3 flétri 3= 3/3 flétri
Senescevol	Foliaire	post-floraison	$\sum \frac{(IF_i + IF_i)}{...}$ avec i le stade donné (F1, F1+20j ou F1+40j)
LAD	Foliaire	post-floraison	$\sum_i \frac{...}{...}$ avec i le stade donné (F1, F1+20j, F1+40j)
Huile	Huile	récolte	Teneur en huile des grains %
RDT	Productivité	récolte	Rendement en grains (quintaux/ha) à 0% d'humidité
RDTH	Productivité	récolte	Rendement en huile (quintaux huile/ha) à 0% d'humidité, RDTH=RDT * Huile
PMG	Productivité	récolte	Poids de mille grains (g)
Nbgrains	Productivité	récolte	Nombre de grains par parcelle Nbgrains=(RDT * Surface parcellaire * 1000)/PMG

Table II.2 : Descriptif des variables phénotypiques mesurées et calculées. Toutes les notations phénologiques correspondent aux nombres de jours après semis lorsque 50% des plantes atteignent le stade en question.

retrouve le rendement (grains et huile) et certaines de ses composantes (poids de mille grain, teneur en huile, teneur en eau), la phénologie (stades de floraison - F1, le début de remplissage du grain -M0 et la maturité physiologique - M3, cf. Figure II.3) et la hauteur de plante. Plusieurs variables ont été calculées à partir de ces mesures, telles que les durées des phases F1-M0 et M0-M3 qui rythment le remplissage du grain, le rendement en huile et le nombre de grains.

Par ailleurs, plusieurs caractères en lien avec la réponse au stress hydrique ont été mesurés :

- Des notations de flétrissement, lorsqu'il était visible, ont été effectuées dans quatre environnements entre la floraison et le stade M3 (trois notations pour CO08\_I et CO08\_NI, et deux notations pour CO09\_I et CO09\_NI).
- Le LAI (Leaf Area Index) est défini comme la surface de feuille photosynthétiquement active par unité de surface de sol. Le LAI est un facteur déterminant pour l'interception lumineuse et pour tous les processus qui en découlent aboutissant à l'allocation des ressources aux grains et donc au rendement. Cette variable est d'ailleurs incluse comme variable d'état dans le modèle écophysique de croissance du tournesol SUNFLO (Casadebaig *et al.*, 2008), dont l'utilisation sera décrite dans la section II.3. Le LAI contrôle la demande en eau à travers l'évapotranspiration qui augmente avec la surface foliaire. En général maximum à la floraison, le LAI diminue ensuite avec la sénescence progressive des feuilles. Lorsque qu'il y présence d'un stress hydrique, certaines plantes réduisent leur surface foliaire pour diminuer les pertes par transpiration (Mitchell *et al.*, 1998). Le LAI est donc un bon indicateur de stress précoce. Dans cet étude il a été mesuré dans neuf environnements à trois dates : à la floraison, 20 jours après la floraison et 40 jours après la floraison, en utilisant le SunScan (DeltaT), appareil qui mesure le rayonnement incident et transmis par le couvert végétal et calcule automatiquement le LAI (cf. encadré). La mesure de trois points au cours du cycle a permis de calculer l'intégrale de la courbe de LAI sur la période post-floraison. Cette variable mesurée (LAD : leaf area duration) traduit le maintien de l'activité photosynthétique, essentielle à la remobilisation des assimilés des feuilles vers les organes reproducteurs durant la phase de remplissage du grain.

## Description du fonctionnement du Sunscan



Le Sunscan est composé d'un bras (1) d'environ 1m constitué de 64 capteurs qui mesurent les radiations photosynthétiquement actives (PAR) incidentes (placé au-dessus du feuillage) et transmises (placé au-dessous du feuillage) par la canopée. Un capteur, relié par ondes radio et positionné au milieu de la parcelle permet de séparer la lumière diffuse de la lumière incidente. Ces informations sont prises en compte dans le calcul du LAI qui est instantané et dont les données sont stockée dans le PDA : Personnel digital assistant (3) Le LAI est déduit à partir d'équations basées sur les travaux de Campbell (1985) et Norman et Jarvis (1975) et qui prennent en compte plusieurs paramètres : les valeurs de PAR mesurées, l'angle du soleil avec le zénith mais aussi l'absorption du couvert et l'angle des feuilles, paramètres fournis par l'utilisateur.



- Un autre caractère influencé par la présence de contraintes hydriques est la sénescence, qui est en général accélérée lorsque le stress s'intensifie. Cette variable a été évaluée par une notation visuelle dans 9 environnements, en considérant le pourcentage de plantes encore « vertes », assimilable au % de plante photosynthétiquement actif (échelle de 0 à 9, avec 9 =90% de plante verte). La notation visuelle a l'avantage d'être rapide et donc bien adaptée au phénotype à haut débit. Cette mesure, que nous appelons « indice foliaire » (IF) ayant été réalisée à plusieurs points du cycle (à floraison, puis 20 à 30 jours après floraison et 40 jours après floraison), a permis également de calculer l'intégrale de l'évolution de l'activité photosynthétique au cours du temps pour la période post-floraison. Cette variable, contrairement au LAI, n'est pas dépendante de la taille des feuilles. Elle est comparable au caractère « stay-green », caractère d'intérêt en sélection car considéré comme favorisant le rendement sous contraintes hydriques (Borrell *et al.*, 2000).

#### 1.1.1.1 Analyses statistiques

Chaque variable phénotypique, qui correspond à une combinaison environnement - caractère, a été analysée dans ASReml-R (Butler *et al.*, 2007) en utilisant le modèle mixte suivant :

$$Y_{ijk} = \mu + G_i + b_j + c_{k(j)} + \varepsilon_{ijk}$$

où  $Y_{ijk}$  est le phénotype observé pour le  $i$ th genotype dans le  $k$ th sous-bloc du  $j$ th bloc,  $\mu$  est la moyenne générale,  $G_i$  est l'effet génétique du  $i$ th genotype considéré comme effet aléatoire,  $b_j$  est l'effet du  $j$ th bloc,  $c_{k(j)}$  est l'effet du sous-bloc  $k$  hiérarchisé dans le block  $j$  et  $\varepsilon_{ijk}$  est l'erreur résiduelle. Les effets bloc et sous-bloc sont considérés comme fixes.

Les valeurs aberrantes ont été enlevées suite à l'observation des histogrammes (les points très extrêmes sont éliminés). Ce premier modèle, que nous appellerons « naïf » permet d'ajuster les variables phénotypiques en fonction des blocs et sous blocs, grâce à la présence des quatre témoins dans chaque répétition, et d'estimer la variance génotypique dans l'objectif de tester si elle est significativement différente de zéro. Pour ce test, le Z-ratio, ratio de l'estimateur de la variance génétique sur son erreur standard qui est calculé par ASReml-R (Butler *et al.*, 2007) a été utilisé. Cette statistique est asymptotiquement, sous l'hypothèse d'une variance nulle, un mélange  $\frac{1}{2}$ ,  $\frac{1}{2}$  entre une loi de Dirac en 0 et une Gaussienne centrée réduite tronquée



sur les réels positifs (Self and Liang, 1987). Une valeur supérieure à 2 permet de rejeter l'hypothèse nulle pour un risque de 5%.

Même si le dispositif expérimental permet de contrôler une part de la variabilité spatiale, d'autres sources de variabilité persistent. Gilmour (1997) en distingue deux principales : la variabilité entre parcelles pouvant provenir d'hétérogénéités de profondeur de sol, de fertilité, de réserve utile et une variabilité d'origine anthropique (appelée « extraneous » par l'auteur) liée principalement aux opérations de conduites de cultures. Par exemple, selon que la dernière opération culturale conduite avant le semis et le semis lui-même sont effectuées dans la même direction ou non, l'hétérogénéité de levée ne sera pas affectée de la même façon. Sont également souvent observés des effets « trains de semoir » engendrés par de petites différences de profondeur de semis. Ces effets sont importants chez le tournesol, pour lequel la qualité de l'implantation est primordiale pour la réussite de la culture (source CETIOM). Par ailleurs, l'utilisation de modèles spatiaux permet de prendre en compte le fait que les parcelles voisines partageant le même micro-environnement, les observations ont plus de chance d'être corrélées.

Nous avons donc choisi de complexifier le modèle naïf en y ajoutant les effets aléatoires des rangées et des colonnes (cf. effet « train de semoir » évoqué ci-dessus) dans un premier modèle spatial (désigné par « RowCol ») afin de modéliser les variations globales ou d'origine anthropique, chaque génotype étant identifié par ses coordonnées (index de rangées et de colonnes du champ). Dans un second modèle spatial (désigné par « ar1 x ar1 »), l'erreur résiduelle a été modélisée par un processus autoregressif d'ordre 1, c'est-à-dire comme le produit d'un modèle de corrélations à travers les rangées avec un modèle de corrélations à travers les colonnes. Ce second modèle est plus adapté à la prise en compte de corrélations environnementales entre parcelles voisines.

La comparaison de ces trois modèles (« naïf », « RowCol », « ar1 x ar1 ») a été menée en utilisant le critère AIC (Aikake information criterion) recommandé pour la prédiction (Yang *et al.*, 2005). Le meilleur modèle a ensuite été utilisé pour extraire les BLUP (Best Linear Unbiased Prediction) pour chaque variable, qui seront utilisés comme variable phénotypique soumise à l'étude d'association.

L'héritabilité au sens large a été calculée à partir du modèle naïf en utilisant la formule

s suivante :  $h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \frac{\sigma_\varepsilon^2}{r}}$ , où  $\sigma_g^2$  est la variance génétique,  $\sigma_\varepsilon^2$ , la variance résiduelle and  $r$ , le

nombre moyen de répétitions par génotype.

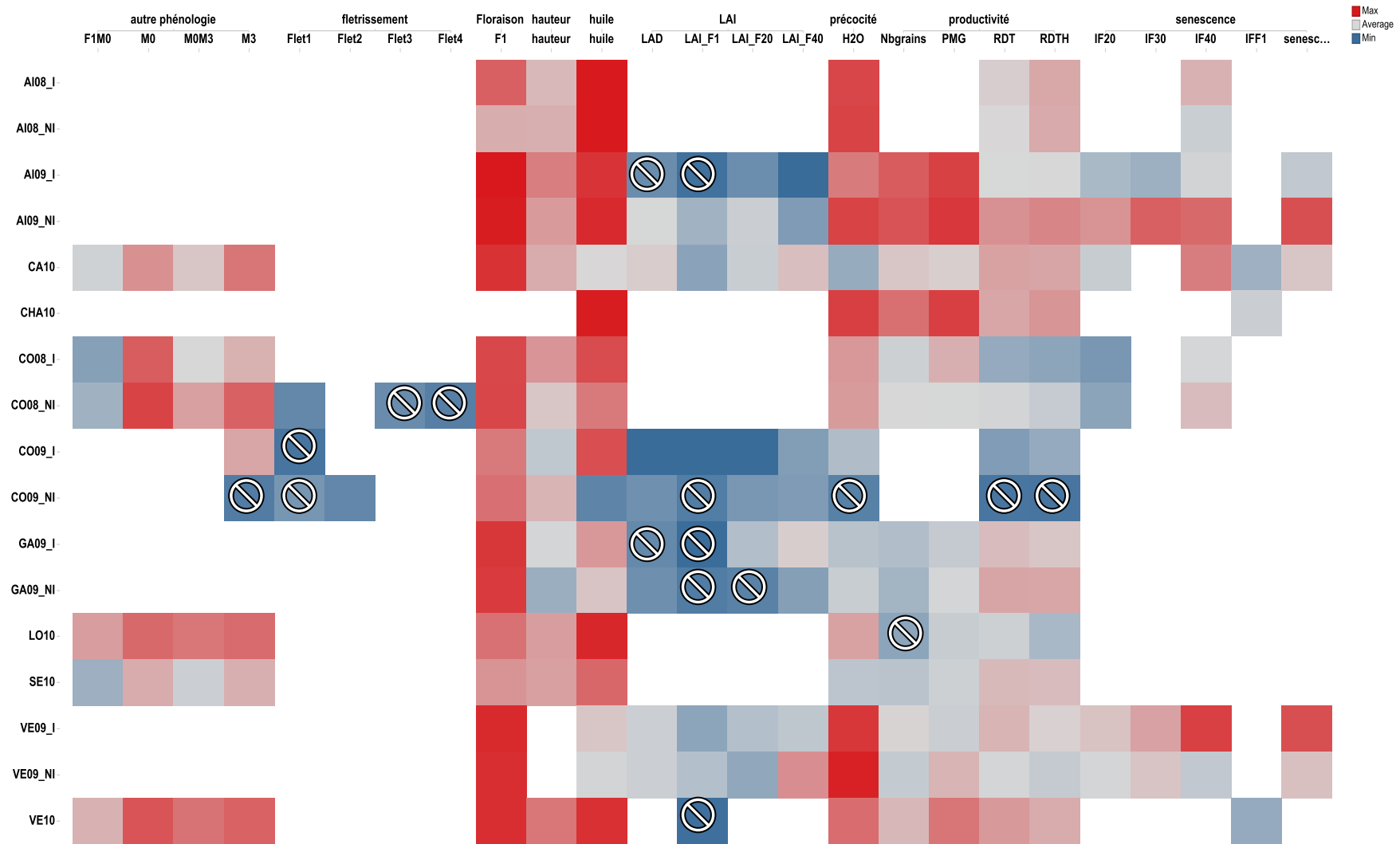



Figure II.4 : Heatmap des héritabilités par caractère et par environnement. Le gradient de couleur est proportionnel aux valeurs d'héritabilités (du bleu au rouge = du minimum au maximum). Les cellules contenant le symbole  correspondent aux variables non significatives.

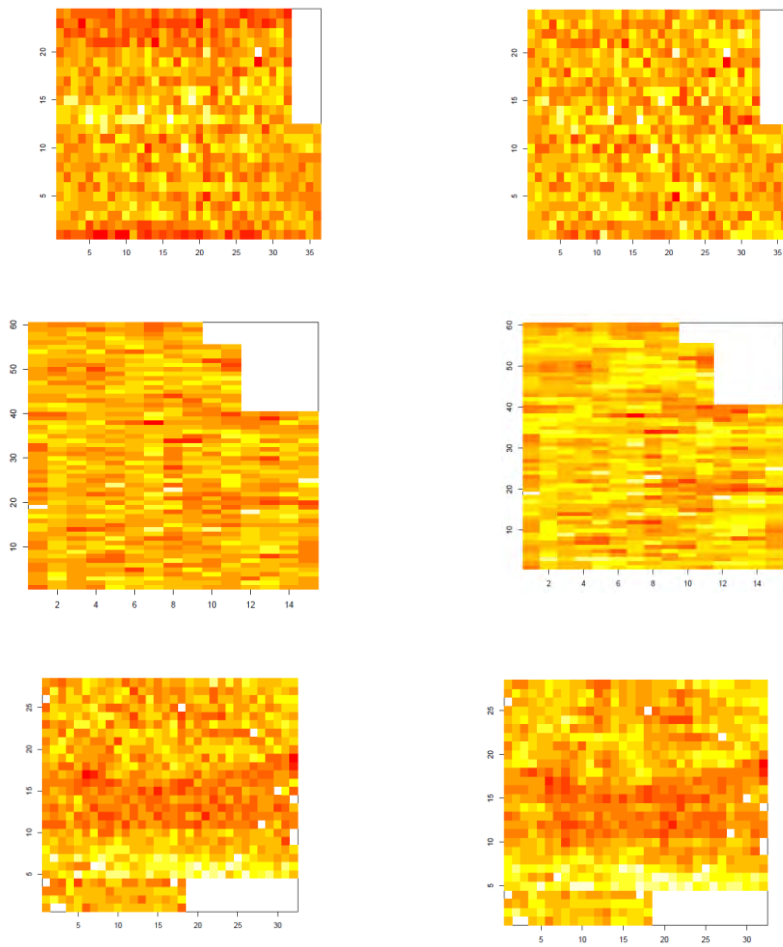
Les corrélations phénotypiques entre variables ont été calculées avec le coefficient de Pearson dans R. Une analyse en composante principale a été menée sur la même matrice de BLUP afin de visualiser l'ensemble des corrélations.

## II.2.2 Résultats et discussions

### II.2.2.1 Ajustement des données phénotypiques avec le modèle naïf

Concernant les effets fixes, la plupart des effets « blocs » sont significatifs alors que seuls 50% des effets « sous-bloc » le sont (Wald test avec Bonferroni à 5%), probablement dû au fait que la variabilité d'un sous-bloc à un autre est minime pour certaines variables. La figure II.4 présente les valeurs d'héritabilité (gradient d'intensité vers les valeurs les plus fortes : du bleu au rouge) pour chaque caractère mesuré dans chacun des 17 environnements. On notera que certaines variables, comme la précocité de floraison, la hauteur, la teneur en huile, l'humidité à la récolte et le rendement sont mesurées dans quasiment tous les lieux. Au contraire, les mesures de LAI ou de PMG (et donc de nombre de grains) sont assez lourdes à mettre en œuvre et n'ont pas pu être acquises dans tous les lieux. Quant au flétrissement, ce type de symptômes n'a été que rarement observé sur le réseau expérimental et n'a donc été noté que dans 3 environnements.

Les variables dont l'effet génotype n'est pas significatif sont indiquées par un symbole. Sur les 222 variables mesurées ou calculées (Table II.2), seules 19 ont une variance génétique non significative ( $Z\text{-ratio} < 2$ ) correspondant pour la plupart à des mesures de LAI ou de flétrissement. Ces deux caractères sont en effet très dépendants des micro-variations (au sein de la plateforme d'essai) de l'environnement. La variabilité spatiale observée sur les données brutes de flétrissement se superpose bien à la variabilité spatiale du rendement, les zones les plus flétries ayant le rendement le plus faible. Le flétrissement traduit donc probablement la capacité de prélèvement en eau nécessaire au rendement et qui diffère selon les zones du sol. De plus, c'est un caractère discret qui, avec seulement 3 notes possibles, nécessiterait probablement un autre traitement statistique. Quant au LAI, il est mesuré avec un appareil (SunScan, Delta-T Devices Ltd, Cambridge, UK) sensible à l'environnement immédiat dont l'angle du soleil par rapport au zénith et la lumière incidente dépendante des conditions météorologiques. Pour mesurer un essai entier, plusieurs heures sont nécessaires, ce qui entraîne des conditions de mesure différentes. De plus le positionnement du bras à capteurs (cf. encadré) est également une source d'erreur. Cette variable est donc peu héritable (0.28 en moyenne).



**a**

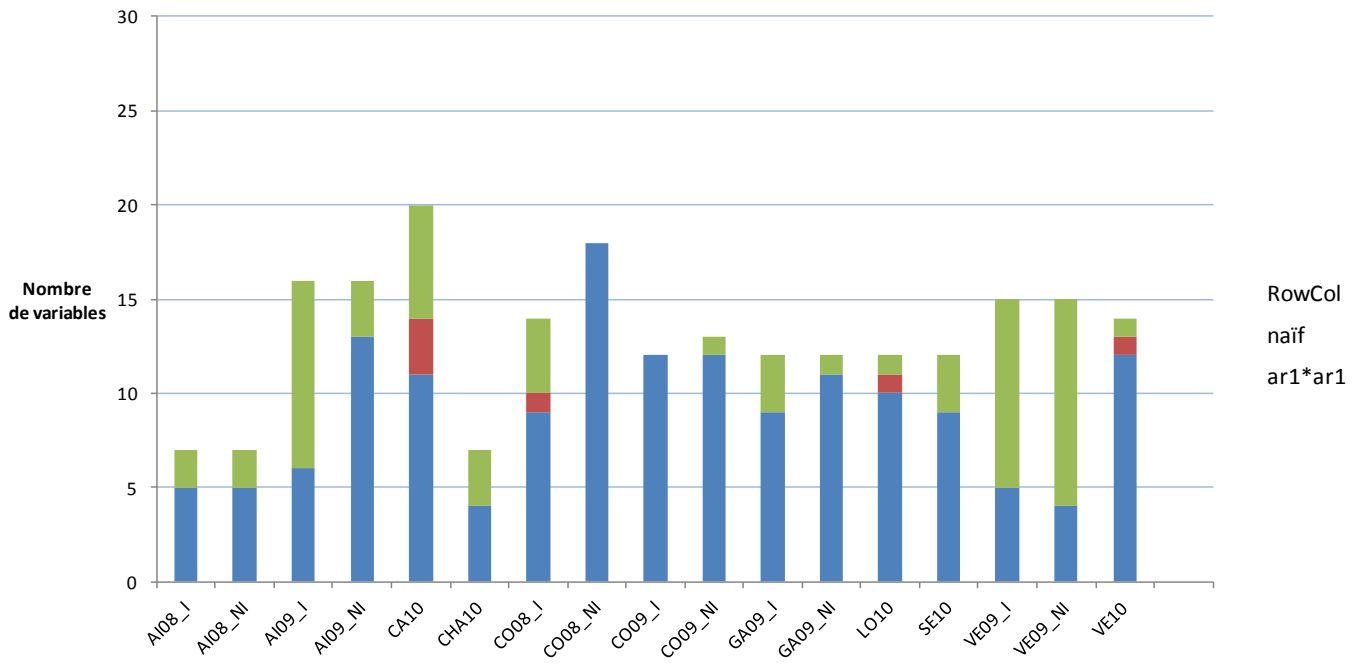
**b**

**Figure II.5 : Cartographie des résidus. a) du modèle naïf b) du meilleur modèle spatial, du haut vers le bas : LAD VE09\_I, Teneur en huile CO09\_NI, IF40\_AI09\_I**

Les héritabilités moyennes diffèrent entre les environnements. CO09\_NI et CO09\_I présentent des héritabilités plus faibles que les autres environnements (respectivement 0.23 et 0.34 en moyenne alors que CHA10 et AI08\_I atteignent des valeurs de 0.77 et 0.74 respectivement). L'environnement CO09\_NI se distingue par le fait que les effets génétiques ne sont statistiquement significatifs ni pour le rendement ni pour la teneur en eau ou en huile. La cartographie des résidus du modèle naïf (Figure II.5) indique que les parcelles voisines en ligne sont très corrélées sur cet environnement, ce qui s'apparente à un effet « train de semoir » (décrit plus haut). Quant à CO09\_I, on observe également la présence de zones, avec des parcelles voisines corrélées pour le rendement, ce qui explique la faiblesse de son héritabilité. C'est aussi le cas pour CO08\_I et CO08\_NI.

Pour AI09\_I, on note la présence de zones pour les caractères d'indice foliaire, de sénescence, qui se superposent aux zones liées au rendement. Il existe probablement une corrélation environnementale forte entre ces caractères, le maintien tardif de l'indice foliaire permettant d'obtenir un meilleur rendement. Enfin, on observe également des zones pour le LAI\_F1 sur GA09\_I et pour le LAD sur VE09\_I.

Les corrélations spatiales dans les résidus permettent donc d'expliquer en partie pourquoi certaines variables ont une héritabilité faible. Globalement, ces corrélations restent peu fréquentes, dû probablement au dispositif expérimental choisi qui permet de corriger une bonne part de la variabilité grâce aux témoins (effet bloc) et aux répétitions des lignées du panel (effet sous-bloc). L'utilisation d'un effet aléatoire pour le génotype dans le modèle statistique permet de prendre en compte la corrélation entre répétitions d'un même génotype. De plus, il a été montré que les BLUP permettaient d'avoir une meilleure prédiction de la vraie valeur génétique (Piepho *et al.*, 2008), notamment dans le cas de dispositifs déséquilibrés. Ce dispositif expérimental offre donc l'avantage de tester un grand nombre de génotypes (de l'ordre de plusieurs centaines à plusieurs milliers), et est bien adapté à des quantités de semences limitées. Cependant la taille des blocs est assez grande et le micro-environnement risque de ne pas être finement corrigé. Il est donc nécessaire de prendre en compte ces variations environnementales à l'aide de modèles spatiaux, afin de révéler au mieux les différents comportements génotypiques et améliorer l'héritabilité des caractères avant la détection de QTL (Moreau *et al.*, 1999).



**Figure II.6 : Distribution des meilleurs modèles pour chaque environnement**



### *II.2.2.2 Ajustement des données phénotypiques avec les modèles spatiaux*

Trois modèles spatiaux ont été comparés en utilisant le critère AIC. 70% des variables ont pour meilleur modèle le modèle autorégressif « ar1 x ar1 » qui modélise l'erreur par le produit des corrélations entre parcelles voisines. Pour la plupart des autres variables, le modèle « RowCol » ressort comme le meilleur. Seules 6 variables sont associées au modèle naïf. La correction spatiale semble apporter une amélioration dans les modèles d'ajustement phénotypiques, d'après ce critère. Il n'y a pas de tendance particulière selon les caractères; par contre certains lieux ont une proportion plus importante de variables ayant l'un ou l'autre des modèles spatiaux comme meilleur modèle (Figure II.6). VE09 est un lieu pour lequel le modèle « RowCol » ressort souvent comme meilleur modèle.

La figure II.5, présenté précédemment, montre la cartographie des résidus après ajustement avec le meilleur modèle spatial. Les effets de bordure pour le LAD à VE09\_I sont bien pris en compte dans le modèle « RowCol ». Par contre, les effets « trains de semoir » ne sont bien expliqués par aucun modèle spatial.

Les prédictions génétiques (BLUP) ont été extraites pour chacun des modèles spatiaux et les corrélations entre BLUP de chaque modèle ont été calculées pour chaque variable. Elles sont en général très fortes, ce qui laisse à croire que les modèles spatiaux n'ont finalement pas apporté beaucoup d'améliorations. Les variables ayant les corrélations les plus fortes sont liées à la phénologie tandis que le flétrissement, le LAI et le rendement présentent les corrélations les plus faibles, donc un potentiel d'amélioration par le modèle spatial plus important. Les résultats détaillés de ces comparaisons de modèles sont disponibles en rapport annexe. Pour la suite des analyses, les prédictions génétiques issues du meilleur modèle ont été conservées.

### *II.2.2.3 Intérêt des caractères choisis*

Avec 14 caractères mesurés et 6 calculés, sur un ensemble de 17 environnements, la quantité de données disponibles dans cette étude rend leur utilisation complexe, notamment car beaucoup de ces caractères sont liés. L'objectif de ce paragraphe est d'identifier l'intérêt des caractères mesurés dans cette étude dans une perspective d'amélioration génétique de la productivité sous contraintes hydriques, en utilisant les résultats d'héritabilité et de corrélations phénotypiques entre variables. Quelques résultats de caractérisation des environnements (section II.3) sont utilisés dans cette partie afin d'aider à l'interprétation. La



Figure II.7 : Boîtes de dispersion des héritabilités du modèle naïf pour chaque environnement.

	AI08_I	AI08_NI	AI09_I	AI09_NI	CA10	CHA10	CO08_I	CO08_NI	CO09_I	GA09_I	GA09_NI	LO10	SE10	VE09_I	VE09_NI	VE10
F1	0.039	-0.012	0.041	0.017	-0.018	-	0.264	0.298	0.101	0.024	0.060	-0.067	0.209	0.351	0.239	-
hauteur	0.100	0.200	0.177	0.120	0.213	-	0.152	0.260	0.116	0.262	0.208	0.258	0.326	-	-	-
H2O	0.099	0.053	0.197	0.060	0.107	0.114	-0.023	0.249	0.041	-0.022	0.005	0.206	0.212	0.211	0.342	0.213
huile	0.014	-0.001	0.088	0.019	0.265	0.039	-0.004	-0.212	0.079	0.167	0.223	0.167	0.005	0.043	0.010	-0.057
F1M0	-	-	-	-	0.040	-	-0.053	0.122	-	-	-	0.063	-0.028	-	-	-
MOM3	-	-	-	-	0.475	-	-0.013	-0.045	-	-	-	0.475	0.395	-	-	0.295
Nbgrains	-	-	0.504	0.536	0.852	0.446	0.534	0.641	-	0.887	-	-	0.645	0.501	0.498	0.620
PMG	-	-	0.216	0.181	0.453	0.288	0.011	0.178	-	0.169	-	0.411	0.241	0.133	0.101	0.326
IFF1	-	-	-	-	0.233	0.339	-	-	-	-	-	-	-	-	-	-
IF20	-	-	0.275	0.357	0.479	-	0.047	0.290	-	-	-	-	-	0.414	0.331	-
IF30	-	-	0.344	0.479	-	-	-	-	-	-	-	-	-	0.419	0.292	-
IF40	0.154	0.206	0.506	0.551	0.473	-	0.024	0.219	-	-	-	-	-	0.565	0.283	-
senescevol	-	-	0.428	0.508	0.574	-	-	-	-	-	-	-	-	0.525	0.338	-
LAI_F20	-	-	0.287	0.464	0.531	-	-	-	0.290	0.226	-	-	-	0.263	0.312	-
LAI_F40	-	-	0.446	0.298	0.483	-	-	-	0.257	0.198	0.139	-	-	0.475	0.523	-
LAI_F1	-	-	-	0.360	0.280	-	-	-	0.189	-	-	-	-	0.324	0.185	-
LAD	-	-	-	0.524	0.584	-	-	-	0.301	-	0.252	-	-	0.414	0.396	-

Table II.3 : Corrélations phénotypiques avec le rendement pour chaque environnement

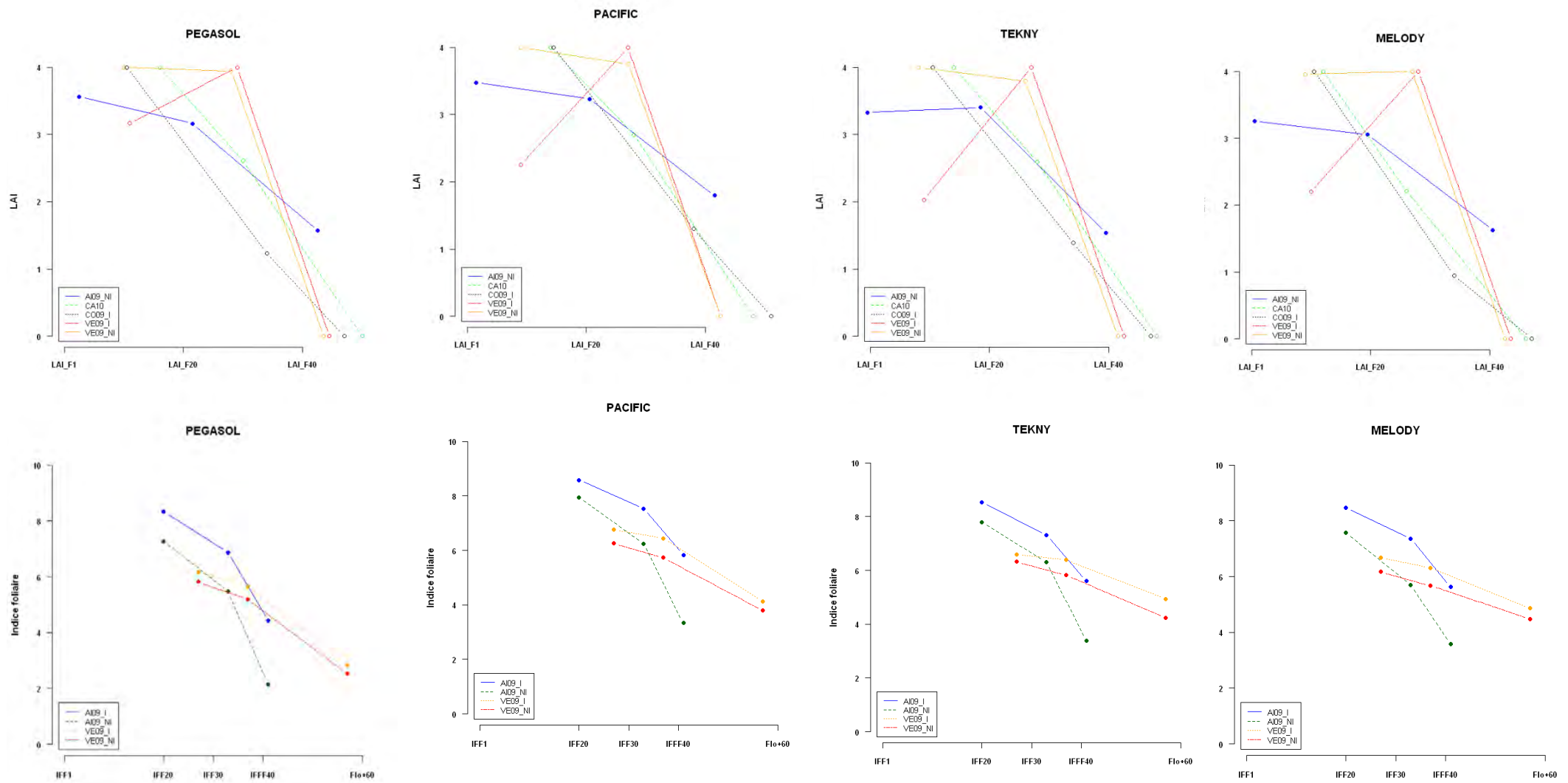
figure II.7 représente les boîtes de dispersion des héritabilités par caractères, la table II.3 présente les corrélations phénotypiques entre le rendement et les autres caractères.

- Phénologie et productivité :

Les caractères de stade phénologique de développement (F1, M0 et M3) ont les héritabilités les plus fortes et sont donc peu influencées par l'environnement. La précocité de floraison (F1) est très corrélée aux autres stades phénologiques, à la hauteur et à l'humidité du grain à la récolte (H2O) dans tous les environnements (Table II.3). Elle est aussi corrélée au rendement grains (RDT) dans 3 lieux : CO08, VE09 et SE10. Dans ces lieux, les géotypes précoces ont moins de grains, le développement de la biomasse a probablement constitué le principal facteur limitant du rendement tandis que pour SE10, le PMG et la durée de remplissage des grains sont aussi corrélés avec le rendement ce qui souligne l'importance de cette phase post-floraison pour ce lieu particulièrement stressant. Ainsi, d'un lieu d'expérimentation à l'autre, différentes composantes de la variabilité génétique s'expriment comme facteur limitant de la productivité, et ceci en fonction des contraintes imposées par les conditions environnementales. Ceci constitue le fondement des interactions géotype - environnement.

Pour les autres lieux où le rendement n'est pas l'avantage des géotypes tardifs, la composante la plus corrélée au rendement sur l'ensemble des lieux est le nombre de grains (Table II.3), alors que le PMG est légèrement moins corrélé voire pas du tout pour CO08. Dans notre étude, la mise en place des grains est donc une étape cruciale dans l'élaboration du rendement, davantage que le remplissage du grain. Nous pouvons remarquer que la teneur en huile n'est pas du tout corrélée au rendement en grains, ces caractères étant le résultat de processus différents.

Concernant les durées de stades phénologiques, la durée semis-floraison (F1M0) présente peu d'intérêt au regard de ses héritabilités assez faibles et de ses corrélations négligeables avec le rendement. Comme déjà démontré chez le tournesol (Chervet *et al.*, 1990), cette variable est indépendante de la durée M0M3. Les héritabilités associées à la durée de stade M0M3 s'étendent de 0.52 à 0.77. Cette variable, qui traduit le temps disponible pour le remplissage des grains, est en général bien corrélée au rendement et présente donc un intérêt pour l'amélioration du rendement.



**Figure II.8 : Evolution du LAI et de l'indice foliaire mesuré à 3 stades du cycle sur 4 génotypes témoins (TEKNY, MELODY, PACIFIC et PEGASOL) dans plusieurs environnements.**

- LAI, indices foliaires et productivité:

Globalement, le LAI (Leaf Area Index) n'apparaît pas comme particulièrement instructif de par sa faible héritabilité (Figure II.4), d'autant plus faible lorsque les notations sont faites au stade de floraison où il est également moins bien corrélé au rendement. Comme déjà mentionné plus haut, l'effet de l'environnement reste marqué pour ce caractère. Par contre à 40 jours après floraison, cette mesure permet de distinguer les génotypes qui maintiennent le plus longtemps leur capacité photosynthétique. L'évolution du LAI le long du cycle est représentée figure II.8 pour les 5 environnements où les mesures sont significatives aux trois points de stades, pour chacun des 4 témoins utilisés (moyenne des parcelles). Contrairement à ce qui était prévu, c'est-à-dire une mesure à floraison, 20 jours et 40 jours après floraison, on observe quelques décalages car les mesures n'ont pas été effectuées en fonction du stade de chaque génotype, mais à trois dates précises déterminées en fonction de la moyenne de floraison du panel. Pour les génotypes témoins par exemple, les trois mesures de LAI ont en réalité été faites à 15 jours puis, de 20 à 30 jours et enfin à plus de 40 jours après floraison. Les points représentés sur les graphes indiquent les vraies dates de mesure. Le LAI mesuré sur chaque hybride du panel peut donc correspondre à un stade un peu différent. L'évolution du LAI et en particulier l'aire sous la courbe (LAD) permet donc d'avoir une mesure plus comparable entre génotypes. La figure II.8 montre des tendances différentes selon l'environnement. AI09\_NI est caractérisé par une décroissance faible du LAI, celui-ci restant élevé à des stades avancés. Cette observation est en concordance avec le fait que cet environnement est très productif et peu stressé. VE09\_I et VE09\_NI, lieux assez productifs, maintiennent leur niveau de LAI jusqu'à environ 30 jours après floraison puis diminuent assez brutalement. Quant à CA10 et CO09\_I, ce sont des lieux stressants pour lesquels le LAI décroît linéairement dès la première notation. On observe par contre très peu de variabilité entre les 4 génotypes témoins. L'évolution du LAI permet donc d'obtenir un indicateur du stress au sein d'un environnement et aussi à l'échelle d'une parcelle, car on observe une variabilité spatiale pour cette mesure souvent corrélée à la variabilité spatiale du rendement (exemple : AI09 ou GA09). Cependant, le LAI semble moins efficace pour discriminer les génotypes.

Comparé au LAI, l'indice foliaire présente une meilleure héritabilité (0.39 à floraison et 0.64 40 jours après floraison) et a l'avantage d'être plus facilement mesurable. Il est également davantage corrélé au rendement (grains et huile) et au nombre de grains que le LAI, notamment dans les lieux à fortes productivité (AI09, VE09, CA10). Le retard de la

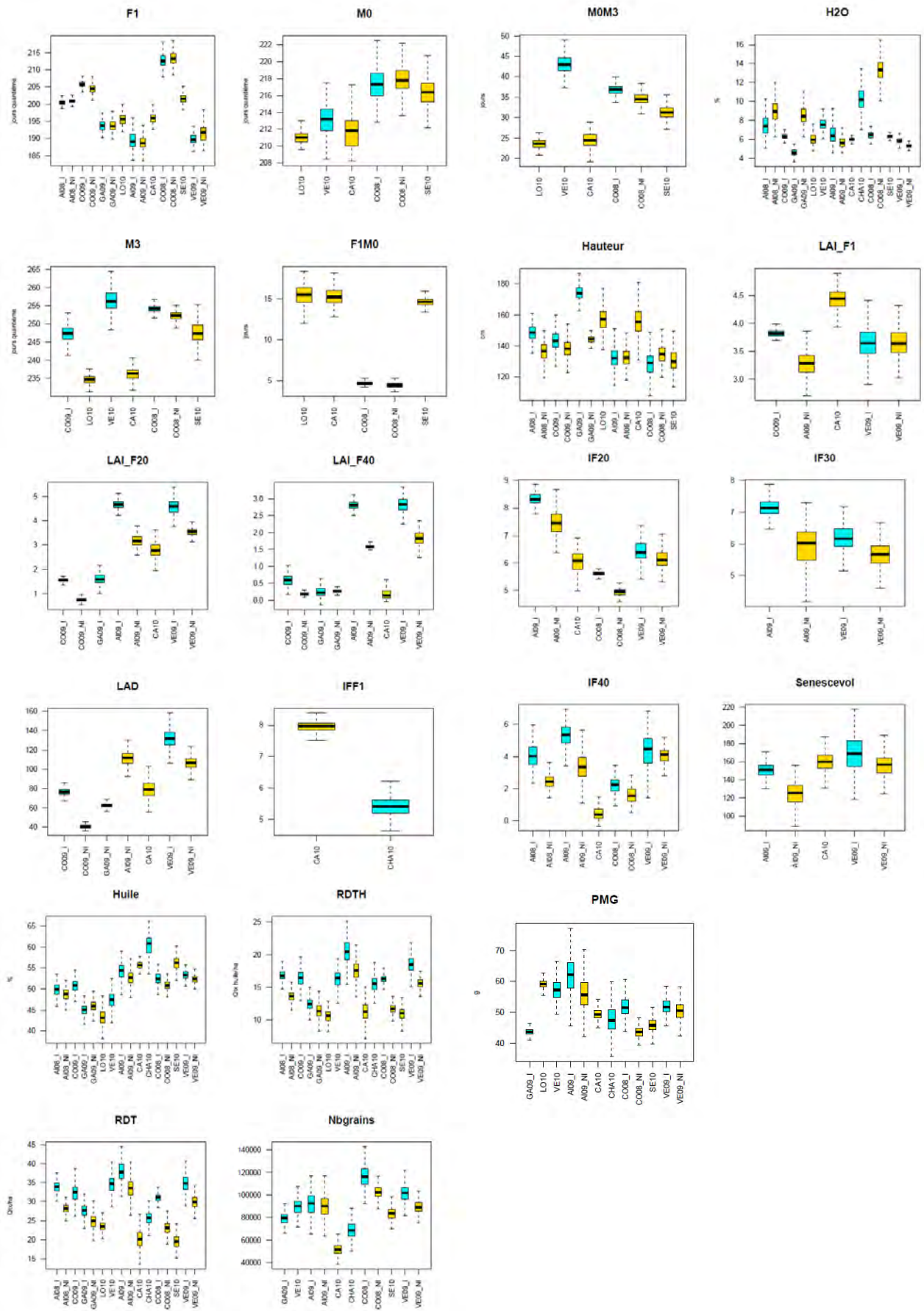


Figure II.9 : Boîtes de dispersion des BLUP sur les différents environnements

sénescence apparaît donc comme une stratégie essentielle pour améliorer le rendement. L'évolution de l'indice foliaire en fonction du stade sur quatre environnements (Figure II.8), montre une différence de pente entre environnements, de la même façon que le LAI. Même si ces pentes sont très similaires entre génotypes, on remarque que l'indice foliaire de PEGASOL est beaucoup plus faible à 40 jours que les autres génotypes. Cet indicateur permet de révéler quelques différences génotypiques, ces génotypes ayant environ la même précocité.

- Comparaison irrigué/non irrigué

En 2008 et 2009, chaque lieu d'expérimentation contenait deux environnements adjacents: l'un irrigué et l'autre non. Les boîtes de dispersion de la figure II.9 permettent de visualiser la différence entre les traitements irrigués ou non sur un même lieu. Le rendement diminue en moyenne de 5 quintaux (15% du rendement irrigué moyen) lorsqu'il n'y pas d'irrigation, cette différence variant de 2.75 (GAI09) à 7.95 quintaux (CO08). Selon les lieux, la différence entre sec et irrigué est donc difficilement contrôlable. Le PMG diminue d'environ 9% et le nombre de grains de 8%. VE09 ne montre pas de diminution de son PMG alors que le nombre de grains diminue. C'est le phénomène inverse pour AIG09. On peut donc parler de phénomènes de compensations entre plusieurs phases du cycle. Les indices foliaires et le LAI sont en général bien impactés par le stress hydrique. Par exemple, l'indice foliaire à 40 jours diminue d'environ 27 % dans les environnements non irrigués par rapport à leurs voisins irrigués. La floraison et la teneur en huile ne montrent pas de différences entre les deux traitements.

De nombreux index de sélection basés sur la différence sec/irrigué ont été proposés dans la littérature. L'index de tolérance au stress (TOL) est la différence entre le rendement en condition irrigué ( $Y_i$ ) et le rendement en condition non irriguée ( $Y_n$ ) (Rosielle et Hambli, 1981). Fischer et Maurer (1978) ont également créé un index (SSI : stress susceptibility index) en considérant le ratio de  $Y_i$  sur  $Y_n$  pour un génotype comparé à ce même ratio sur l'ensemble des génotypes. Chez le tournesol, Darvishzadeh *et al.*, (2011) ont étudié l'héritabilité de 8 index sur 21 génotypes dérivant d'un diallèle incomplet de 6 lignées, et en ont isolé 4 présentant une héritabilité suffisante et donc un intérêt en sélection. Cependant, ces index ne permettent pas toujours de sélectionner des individus qui ont aussi un rendement supérieur en conditions non stressante. Il est donc important de choisir un index en veillant





aussi à ce principe car l'objectif est d'obtenir des variétés performantes dans les lieux à potentiel en limitant les pertes dans les lieux stressés (stabilité).

La mise en place d'expérimentations irrigué/non irrigué présente des contraintes non négligeables. Le contrôle de l'irrigation est parfois complexe, de plus, une année trop pluvieuse peut masquer la différence sec/irrigué. Certains environnements secs se retrouvent au même niveau de potentiel que des environnements irrigués sur d'autres lieux, ces derniers pouvant présenter des niveaux de stress hydriques plus intenses, comme en atteste la difficulté des géotypes à maintenir l'indice foliaire (ex pour CO09\_I, Figure II.8). A partir de 2010, la stratégie d'expérimentation choisie dans le cadre du programme OLEOSOL (d'où sont extraites les données utilisées pour cette thèse) a donc changé de voie, en multipliant les lieux plutôt que d'effectuer des essais irrigués/non irrigués sur chacun des lieux. L'objectif était donc de prendre en compte plus de scénario de stress, représentatifs des zones de production du tournesol ciblés par les entreprises de sélection. L'approche « index de tolérance au stress » n'a donc pas été étudiée en comparant les traitements irrigué et sec d'un même lieu mais a été globalisée à l'ensemble du TPE (Target Population of Environments) et sera décrit dans la section II.3.

## II.3 Analyse multilocale

### II.3.1 Interaction génotype x environnement

#### II.3.1.1 Matériels et méthodes

Afin d'estimer l'importance de l'interaction génotype - environnement (GE) pour chaque caractère, le phénotype a été modélisé selon la formule suivante :

$$Y_{ijkl} = \mu + e_l + G_i + b_{j(l)} + c_{k(jl)} + GE_{il} + \varepsilon_{ijkl}$$

où  $Y_{ijkl}$  est le phénotype du génotype  $i$  mesuré dans la répétition  $k$  du bloc  $j$  de l'environnement  $l$ ,  $\mu$  est la moyenne générale,  $E_l$  est l'effet de l'environnement  $l$ ,  $G_i$  est l'effet du génotype  $i$ ,  $b_{j(l)}$  est l'effet du  $j$ th bloc hiérarchisé dans l'environnement  $l$ ,  $c_{k(jl)}$  est l'effet du sous-bloc  $k$  hiérarchisé dans le bloc  $j$  de l'environnement  $l$ ,  $GE_{il}$  est l'effet de l'interaction entre le génotype  $G_i$  et l'environnement  $E_l$ ,  $\varepsilon_{ijkl}$  est l'erreur résiduelle. Le

Caractère	modèle global						
	n_envi	VG	p-value	VGE	p-value	VGE/VG	VE
F1	16	3.22 (0.25)	0	1.45 (0.05)	0	0.45	1.33 (0.02)
M0	6	3.05 (0.28)	0	2.17 (0.14)	0	0.71	3.11 (0.09)
M3	8	4.59 (0.42)	0	3.20 (0.24)	0	0.70	7.82 (0.20)
Hauteur	14	68.63 (5.6)	0	21.83 (2.00)	0	0.32	113.18 (2.11)
IFF1	3	0.03(0.01)	2.54E-04	0.09 (0.02)	2.49E-14	21.99	0.35 (0.01)
IF20	7	0.16(0.016)	0	0.09 (0.01)	0	40.64	0.40 (0.01)
IF30	4	0.29(0.03)	0	0.009 (0.006)	2.07E-02	0.03	0.34 (0.01)
IF40	9	0.40(0.04)	0	0.27 (0.02 )	0	0.68	0.73 (0.02)
Senecevol	5	208.83 (18.77)	0	68.82 (7.99)	0	0.33	237.46 (6.96)
Nbgrains	14	7.4E+07(6.2E+06)	0	3.44E+07(2.98E+06)	0	0.47	1.38E+08(2.93+06)
H2O	17	1.07 (0.09)	0	0.89 (0.04)	0	0.83	1.80 (0.03)
huile	17	4.09 (0.31)	0	1.00 (0.05)	0	0.24	2.89 (0.05)
RDTH	17	1.43(0.12)	0	0.80 (0.07)	0	0.56	4.37 (0.08)
PMG	14	19.34(1.54)	0	6.74 (0.52)	0	0.35	22.71 (0.49)
RDT	17	3.17(0.29)	0	2.96 (0.22)	0	0.93	14.2 (0.23)
LAI_F1	10	0.04(0.006)	0	0.012(0.007)	1.79E-02	0.30	0.70 (0.01)
LAI_F20	9	0.04(0.005)	0	0.037 (0.007)	1.69E-10	0.99	0.41 (0.01)
LAI_F40	9	0.03(0.003)	0	0.019 (0.003)	1.54E-12	0.60	0.19 (0.004)
LAD	9	43.94(5.10)	0	19.39 (5.4)	9.02E-06	0.44	330.66 (7.07)
F1M0	6	0.05(0.05)	0.13	1.27 (0.11)	0	25.83	2.73 (0.08)
M0M3	6	2.05(0.26)	0	3.79 (0.27)	0	1.84	6.46 (0.19 )

**a**

Caractère	SOLR001M/ AT0521						
	n_envi	VG	p-value	VGE	p-value	VGE/VG	VE
F1	8	4.82 (0.37)	0	1.11 (0.06)	0	0.23	1.5 (0.04)
M0	4	4.78 (0.45)	0	1.72 (0.21)	0	0.36	4.4 (0.16)
M3	4	5.89 (0.59)	0	2.33 (0.33)	0	0.40	7.63 (0.28)
hauteur	6	72.92 (6.37)	0	17.36 (2.64)	5.55E-16	0.24	96.16 (2.7)
IFF1	2	0.07 (0.02)	3.79E-04	0.1 (0.03)	1.95E-06	1.43	0.44 (0.02)
IF20	7	0.16 (0.02)	0	0.09 (0.01)	0	0.56	0.4 (0.01)
IF30	4	0.29 (0.02)	0	0.01 (0.01)	0.02	0.03	0.34 (0.01)
IF40	7	0.45 (0.04)	0	0.22 (0.02)	0	0.49	0.69 (0.02)
senescevol	5	208.83 (18.77)	0	68.82 (7.99)	0	0.33	237.46 (6.95)
Nbgrains	9	9.73E+07(8.26E+06)	0	3.15E+07 (3.11 E+06)	0	0.32	1.37E+08(3.13E+06)
H2O	9	1.57 (0.13)	0	0.69 (0.04)	0	0.44	1.47 (0.03)
huile	9	4.08 (0.32)	0	0.94 (0.07)	0	0.23	2.53 (0.06)
RDTH	9	2.13 (0.19)	0	0.7 (0.08)	0	0.33	3.6 (0.08)
PMG	9	23.93 (1.93)	0	5.98 (0.54)	0	0.25	22.28 (0.51)
RDT	9	6.18 (0.53)	0	1.78 (0.23)	0	0.29	11.08 (0.25)
LAI_F1	5	0.11 (0.01)	0	0.02 (0.01)	0	0.18	0.55 (0.01)
LAI_F20	5	0.08 (0.01)	0	0.05 (0.01)	5.08E-07	0.63	0.46 (0.01)
LAI_F40	5	0.05 (0.005)	0	0.02 (0.006)	5.14E-08	0.40	0.2 (0.01)
LAD	5	88.47 (9.44)	0	21.7 (6.71)	1.57E-05	0.25	286.99 (8.16)
F1M0	4	1.8e-06 (6.5e-08)	0.5	1.24 (0.12)	0	6.89E+05	3.12 (0.11)
M0M3	4	1.99 (0.33)	1.54E-14	3.39 (0.38)	0	1.70	7.92 (0.29)

**b**

**Table II.4 : Variances pour le modèle global (a) et les modèles avec séparation des environnements selon la paire de testeur ; groupe 1 : testeurs SOLR001M/AT0521 (b) (explications plus détaillées dans la suite de la table)**

génotype, l'interaction GE et la résiduelle sont des effets aléatoires. La significativité des composantes de la variance a été estimée à partir de test de rapport de vraisemblance.

Dans ce modèle, l'effet « environnement » représente en réalité la combinaison de l'environnement agronomique et de l'environnement génétique induit par les testeurs utilisés. En effet, les hybrides issus des croisements 83HR4gms/lignées B ou FS71501/lignées R (première paire de testeurs) ont été évalués sur 8 environnements (que nous appellerons groupe 1) et les hybrides issus des croisements SOLR0001M/lignées B ou AT0521/lignées R (seconde paire de testeurs) sur neuf environnements (groupe 2). Cette confusion partielle de l'effet environnement avec l'effet paire de testeurs se répercute aussi dans l'interaction génotype-environnement et peut affecter la part de variance due à cette interaction. Ce modèle fait aussi le postulat de l'homoscédasticité de la variance des effets GE dans les deux groupes. Pour se libérer de la confusion partielle et du postulat, le modèle statistique décrit ci-dessus a été appliqué séparément, dans un second temps, aux deux sous-ensembles d'environnements correspondant à la même paire de testeur : groupe 1 et 2.

Les corrélations entre toutes les variables phénotypiques ajustées (BLUP) ont été calculées dans R en utilisant le coefficient de Pearson. Une analyse en composante principale a été effectuée pour chaque caractère à partir des BLUP sur l'ensemble des environnements où le caractère a été mesuré en utilisant la fonction PCA du package FactoMineR (<http://cran.r-project.org/web/packages/FactoMineR/index.html>). Les individus ayant plus de 80% d'environnements manquants ont été éliminés. Les autres données manquantes ont été inférées en les remplaçant par la moyenne de l'environnement.

### *II.3.1.2 Résultats et discussion*

La table II.4 présente les composantes de la variance estimées à partir du modèle global dans tous les environnements et dans les environnements correspondant à l'une ou l'autre paire de testeur (groupes 1 et 2).

La plupart des caractères présente un effet génotype significatif d'après les tests de rapport de vraisemblance. L'interaction GE est également significative pour tous les caractères sauf quelques mesures de LAI pour le groupe 1. L'analyse par paire de testeurs met en évidence la contribution de l'effet « paire de testeurs » sur la variance de l'interaction GE pour la plupart des caractères, mais cette interaction reste significative au sein de chacun des groupes 1 et 2. L'intervalle de confiance des variances a été calculé pour chacun des groupes. Pour la majorité des variables, il n'y a aucun chevauchement entre les intervalles de confiance, ce qui

Caractère	83HR4gms/FS71501						
	n_envi	VG	p-value	VGE	p-value	VGE/VG	VE
F1	8	2.87 (0.23)	0	0.4 (0.03)	0	0.14	1.14 (0.03)
MO	2	1.1 (0.17)	1.79E-13	1.26 (0.14)	0	1.15	1.07 (0.05)
M3	4	4.29 (0.49)	0	2.82 (0.35)	0	0.66	7.77 (0.27)
hauteur	8	80.79 (6.64)	0	9.32 (2.22)	3.35E-10	0.12	120.8 (2.9)
IFF1	1	-	-	-	-	-	-
IF20	1	-	-	-	-	-	-
IF30	1	-	-	-	-	-	-
IF40	2	0.55 (0.08)	0	0.14 (0.05)	7.12E-12	0.25	0.86 (0.04)
senescevol	1	-	-	-	-	-	-
Nbgrains	5	3.52E+07(5.49E+06)	0	1.76E+07(6.75E+06)	6.28E-05	0.50	1.37E+08 (7.37E+06)
H2O	8	0.92 (0.09)	0	0.67 (0.06)	0	0.73	2.15 (0.05)
huile	8	4.36 (0.35)	0	0.36 (0.08)	2.72E-09	0.08	3.46 (0.09)
RDTH	8	0.98 (0.12)	0	0.44 (0.11)	3.42E-07	0.45	5.13 (0.14)
PMG	5	11.41 (1.34)	0	2.62 (1.21)	1.44E-03	0.23	23.09 (1.3)
RDT	8	2.12 (0.28)	0	2.54 (0.38)	7.77E-16	1.20	16.71 (0.41)
LAI F1	5	0.01 (0.01)	0.04483	3.7e-08 (8.7e-10)	0.5	3.70E-06	0.82 (0.02)
LAI_F20	4	0.02 (0)	5.38E-07	0.001 (0)	3.99E-01	0.05	0.33 (0.01)
LAI_F40	4	0.02 (0)	2.66E-13	0.009 (0)	2.02E-02	0.45	0.17 (0.01)
LAD	4	22.14 (5.53)	2.96E-07	1.6e-05 (4.3e-07)	0.5	7.23E-07	369.76 (10.11)
F1M0	2	0.65 (0.16)	2.62E-05	0.92 (0.18)	9.44E-16	1.42	1.77 (0.1)
MOM3	2	2.39 (0.51)	1.48E-06	3.97 (0.51)	0	1.66	3.92 (0.19)

c

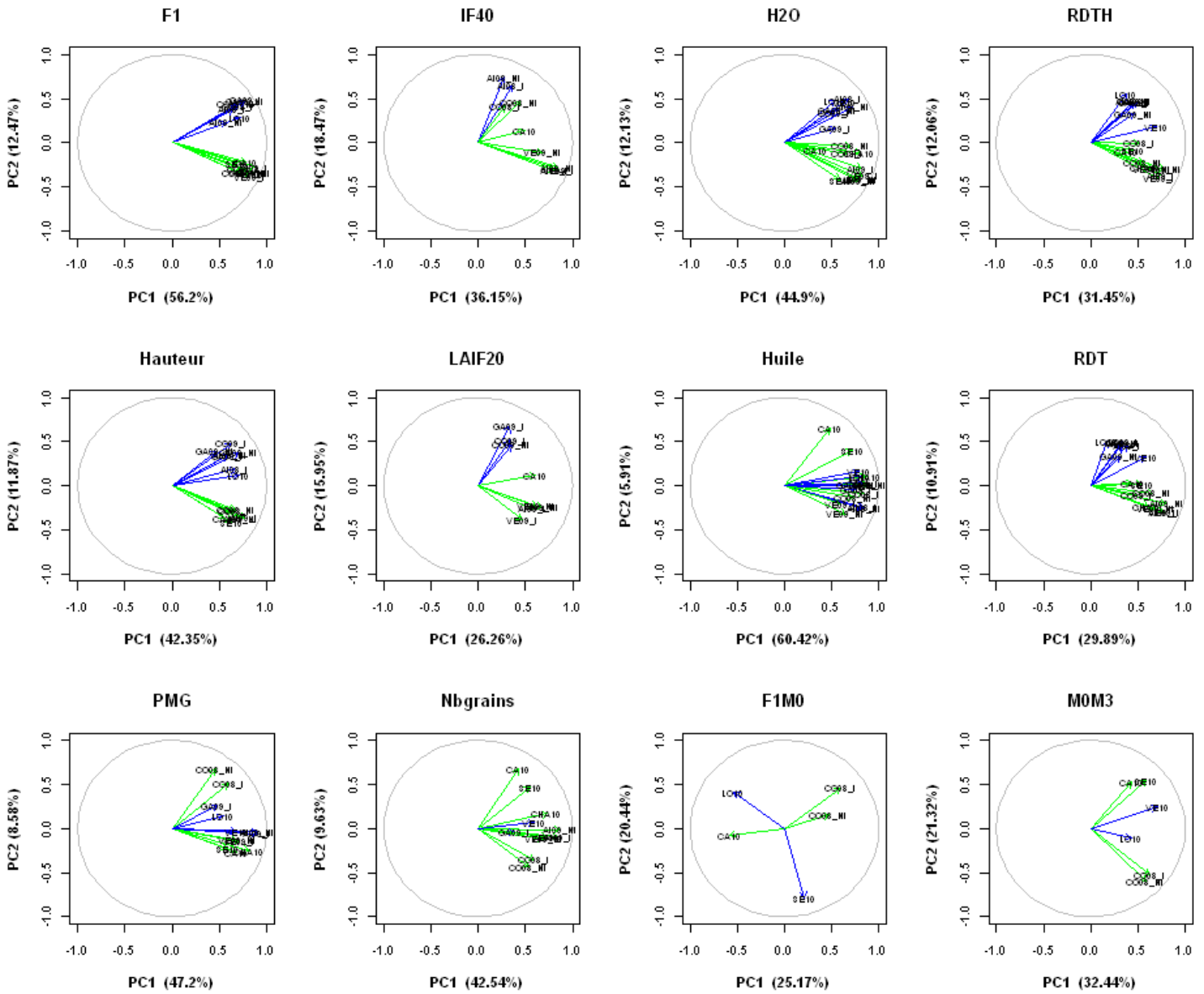
**Table II.4 (suite) : Variances groupe 2 : testeurs 83HR4gms/FS71501(c).** Pour chaque caractère, les variances génétiques (VG), variances de l'interaction génotype x environnement (VGE) et variance de l'erreur (VE) sont présentées avec leur erreur standard entre parenthèses. Les p-values du test de ratio de vraisemblance sont données pour VG et VGE, et le ratio VGE/VE a été calculé. Le nombre d'environnements associés à chaque modèle est indiqué dans la colonne « n\_envi »

prouve que le postulat d'homoscédasticité était incorrect. Lorsque l'on compare les groupes 1 et 2, les variances génétiques et d'interaction GE sont plus importantes pour le groupe 1. Au sein de ce groupe, certains lieux tels que AI09, VE09 et CHA10 sont à potentiel élevé, et montrent davantage de variabilité pour le rendement (Figure II.9).

Il est intéressant d'interpréter les ratios de la variance GE sur la variance génétique afin d'apprécier l'importance de l'interaction GE. La teneur en huile, la hauteur et la date de floraison présentent des ratios assez faibles, alors que la tardiveté à la récolte, mesurée par la teneur en eau des graines, et le rendement se situent dans la gamme de ratio la plus élevée (ratios égaux à 0.83 et 0.93 respectivement). La part de variance GE pour les caractères de tardiveté et de rendement est donc importante par rapport à la variance génétique. Les boxplot des héritabilités par caractère (précédemment, Figure II.8) montraient une forte variabilité pour la tardiveté (de 0.11 à 0.94), ce qui illustre bien que l'impact de l'environnement sur la tardiveté est plus ou moins marqué, notamment car les essais peuvent être récoltés à des stades variables. Le rendement présente également des héritabilités faibles sur quelques environnements. Au contraire, les caractères présentant une faible interaction GE sont ceux ayant les héritabilités figurant parmi les plus fortes sur le plus grand nombre d'environnements. Le rendement en huile et les composantes du rendement présentent des interactions GE intermédiaires.

Afin de comprendre la nature de l'interaction GE, nous avons réalisé des ACP (Figure II.10) sur les matrices de BLUP (individus x environnements) et moyenné les corrélations phénotypiques entre les environnements pour une même variable (Figure II.11).

Dans la figure II.10, la plupart des caractères sont représentés dans le premier plan de l'ACP; ceux n'étant pas représentés, tels que certaines mesures d'indice foliaire ou de LAI, ont des profils très similaires aux autres stades du cycle figurés. Sur l'ensemble des ACP, la variance de la première composante principale s'étend de 25% pour la durée F1M0 à 60% pour la teneur en huile. Cette première composante témoigne d'un niveau relativement élevé de la corrélation entre les environnements. La teneur en huile, tout comme la floraison apparaissent donc comme fortement corrélées entre les environnements. Dans la figure II.11 a, les barres horizontales, représentant la médiane et le quartile à 75%, permettent de classer les variables selon leur niveau de corrélation en plusieurs paliers. On retrouve à nouveau la floraison et la teneur en huile sur le plus haut palier (corrélations moyennes supérieures à 0.50 environ) ainsi que l'indice foliaire à 30 jours. La tardiveté, la hauteur, le stade M0 et le PMG côtoient le deuxième palier (corrélations moyennes entre 0.10 et 0.30). Enfin le dernier palier, regroupe



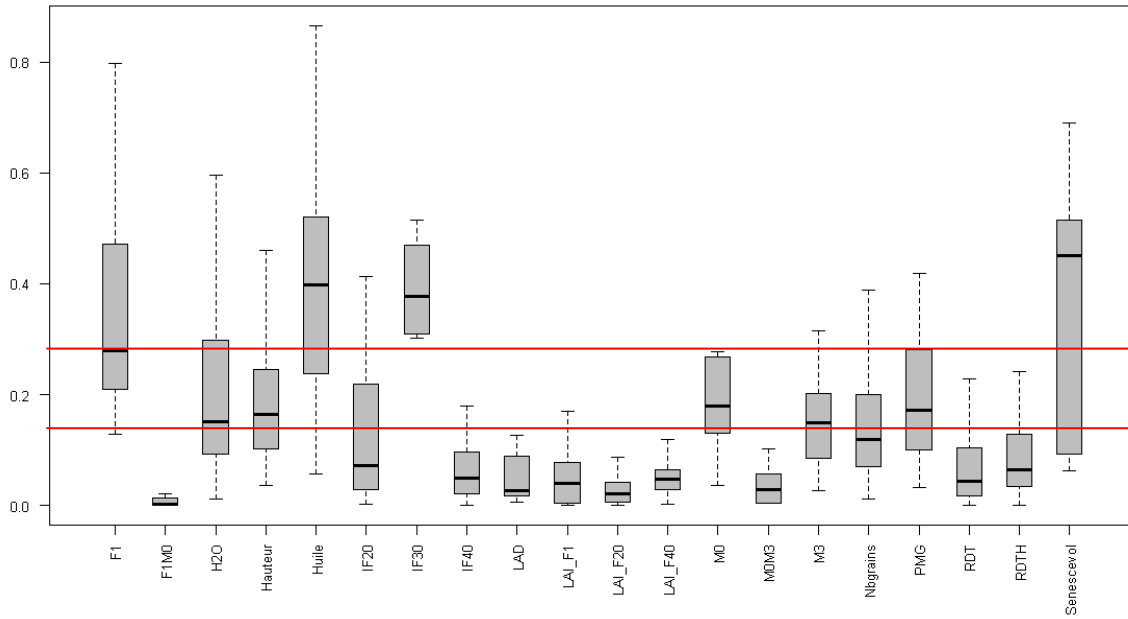
**Figure II.10: Analyses en composante principale à partir des BLUP pour quelques caractères.**  
 Les flèches vertes correspondent aux environnements du groupe 1 (paire de testeur « SOLR0001M/AT0521 »), les flèches bleues correspondent aux environnements du groupe 2 (paire de testeurs « 83HR4gms ou FS71501 »).

50% des données avec des caractères ayant des corrélations la plupart du temps inférieures à 0.10 et donc l'interaction GE la plus forte : LAI, F1M0, M0M3 et le rendement (en grains et en huile).

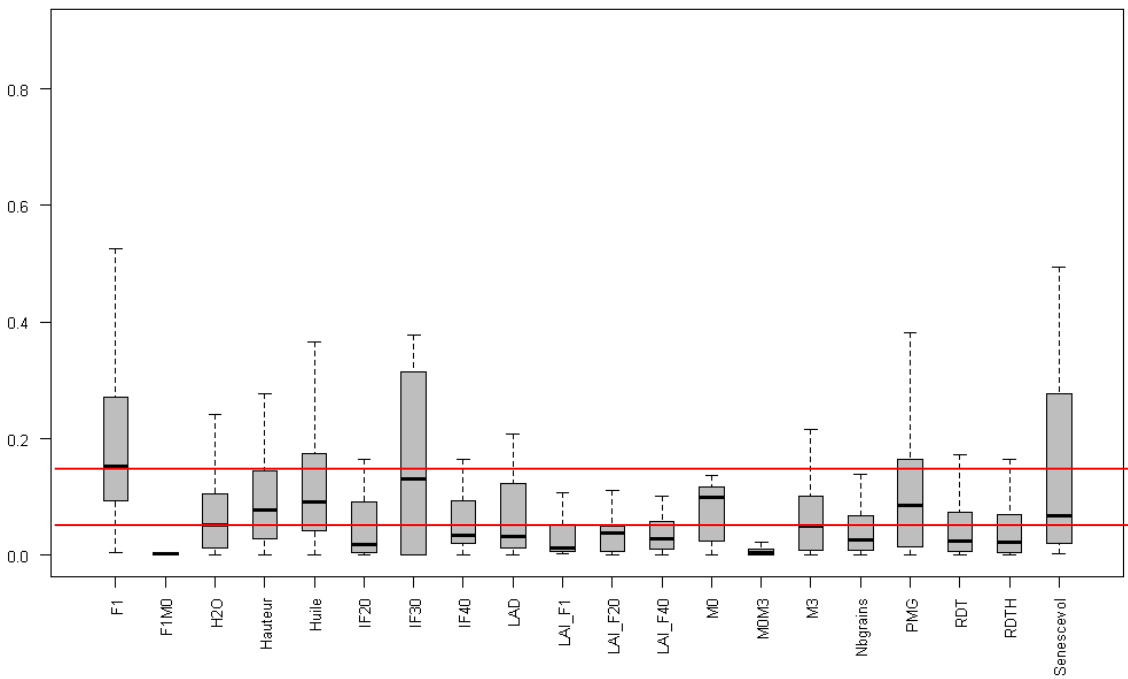
Dans la figure II.10, la deuxième composante principale (PC2) traduit pour la plupart des caractères une structuration des environnements basée sur la paire de testeurs utilisée (flèches bleues ou vertes). Cette composante n'est que de 5% pour la teneur en huile, caractère peu marqué par l'effet du testeur. Pour la floraison, la précocité, le rendement et la hauteur, la PC2 est d'environ 12% et sépare bien les deux groupes de « paire de testeur ». Chaque testeur semble donc apporter une contribution génétique importante qui entraîne des corrélations phénotypiques plus fortes entre environnements ayant les mêmes testeurs. La moyenne des rendements pour chaque groupe de testeurs est cependant similaire, indiquant que les deux testeurs ont une productivité équivalente. Il est à noter qu'aucune information n'est disponible sur les testeurs (distances génétiques par exemple).

Le nombre de grains et le poids de mille grains présentent cependant un comportement légèrement différent. Pour le PMG, malgré un regroupement des environnements sur la base de la paire de testeur, CO08\_NI et CO08\_I se distinguent en se rapprochant des environnements ayant l'autre paire de testeurs. Pour le stade M0M3, ils sont également à l'opposé des environnements du même groupe. Comme mentionné dans la partie « Intérêts des caractères choisis », ce lieu présente un profil atypique. Enfin, pour le caractère IF40, il n'y a pas de distinction entre les paires de testeurs : les lieux CO08 et AI08 sont davantage corrélés même si ils ont des testeurs différents.

On peut donc conclure que la paire de testeur utilisée structure la variabilité phénotypique mais ne constitue pas un effet majeur car la première composante principale, qui explique la plus grande part de variabilité phénotypique, montre des corrélations importantes entre lieux. Pourtant l'interaction GE est bien présente, et également au sein de chaque groupe d'environnements. L'effet testeur, confondu avec le type de lignée (B ou R) n'a pas pu être testé, mais on peut supposer qu'il n'est pas négligeable, car c'est bien la partition majeure du panel qui apparaît à travers les données moléculaires (cf. Chapitre III). Nous faisons donc l'hypothèse que les corrélations phénotypiques observées entre environnements sont en partie dû au fait que les lignées B donnent des hybrides phénotypiquement plus proches entre eux qu'avec ceux des lignées R. Cette structuration, de la même manière que lorsqu'elle cause des fréquences alléliques différentes entre groupes et biaise le déséquilibre de liaison, peut également biaiser les corrélations phénotypiques entre lieux. Pour vérifier cette hypothèse,



**a**



**b**

**Figure II.11 : Distribution des corrélations phénotypiques entre environnements. Les lignes rouges séparent les quantiles (25, 50 et 75%), basés sur l'ensemble de la distribution des corrélations**



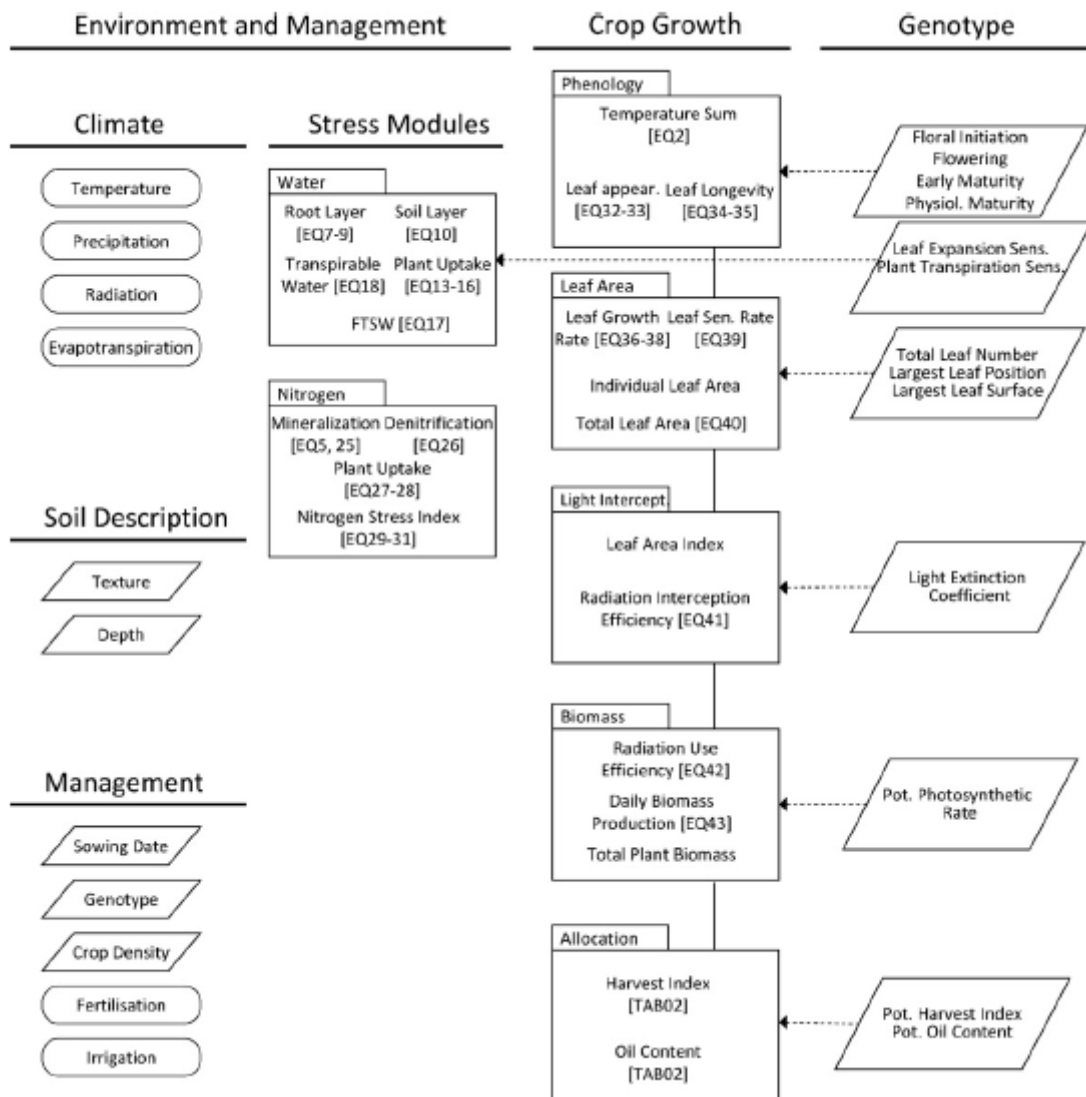
nous avons appliqué la même méthode que celle utilisée dans l'article de Mangin *et al.*, (2011) pour corriger le déséquilibre de liaison de l'appareillement et la structure. Le coefficient de corrélation de Person  $r^2$  a été décomposé de manière à en extraire la partie dépendante de la structure ; le nouveau coefficient de corrélation obtenu entre deux environnements est nul si les deux environnements sont indépendants. La figure II.11 b représente la distribution des corrélations corrigées par la structuration. On remarque que toutes les corrélations sont désormais inférieures à 0.20, à part la floraison et l'indice foliaire qui maintiennent quelques valeurs au-dessus de ce seuil. L'interaction GE est donc très significative dans ce réseau d'essais.

En conclusion, nous pouvons définir quelques tendances. Les corrélations entre environnements irrigués et sec d'un même lieu sont en général plus importantes qu'entre lieux différents. Ceci confirme à nouveau que le traitement irrigué/sec n'a que peu d'influence sur la réponse différentielle des géotypes, et que si l'on cherche à minimiser l'interaction GE dans le cadre par exemple de la sélection de variétés performantes, il faut multiplier le nombre de lieux.

Une stratégie souvent évoquée dans la littérature pour diminuer l'interaction GE est de regrouper les lieux similaires et de travailler sur la moyenne des variables. Dans notre étude, à part AI09 et VE09, lieux à fort potentiel et très corrélés, le regroupement possible d'environnements n'apparaît pas clairement. Les caractères d'indice foliaire et de LAI ne permettent pas de différencier des groupes de lieux, d'autant plus que ces données sont très déséquilibrées. Il est donc nécessaire de continuer à caractériser les environnements, et c'est pour cela que nous avons utilisé un modèle écophysique basé sur le comportement de variétés utilisées comme témoins dans les essais.

### II.3.2 Description du modèle SUNFLO

Le modèle SUNFLO est un modèle de culture développé par Casadebaig *et al.*, (2008) pour simuler de manière dynamique la croissance de variétés de tournesol en interaction avec les différentes composantes de leur environnement abiotique. Par rapport à d'autres modèles de culture développés pour le tournesol (Chapman *et al.*, 1993 ; Villalobos *et al.*, 1996 ; Oereyra-Irujo *et al.*, 2007), ce modèle a fait le choix de s'appuyer sur un nombre limité de paramètres, ce qui permet notamment de pouvoir simuler les comportements de nouveaux géotypes à une fréquence compatible avec le flux d'innovation variétale.



**Figure II.12 : Représentation des différents modules du modèle SUNFLO (Casadebaig et al., 2010).** Le module de croissance au centre dépend du module « Genotype » et du module « Environment and Management ». (Environnement et conduite de culture). Les parallélogrammes représentent les paramètres (c'est-à-dire les constantes) et les ellipses les variables d'entrée qui peuvent varier chaque jour. Les rectangles sont les variables intermédiaires avec les équations dont le détail est accessible dans la publication citée.

Actuellement, il est utilisé par le Groupe d'Etude et de contrôle des Variétés et des Semences (GEVES) pour caractériser les réseaux d'essais lors des tests d'inscription des variétés au catalogue par le Comité Technique Permanent de la Sélection (CTPS), et également en post-inscription par le Centre Technique Interprofessionnel des Oléagineux et du Chanvre (CETIOM) où il permet de fournir des avis concernant la conduite de culture (ex : densité et dates de semis). Il est également utilisé par les sélectionneurs. Un des objectifs à terme est d'aider à choisir dans des conditions de milieu données, les meilleures combinaisons variété-itinéraire technique.

Dans ce modèle (décrit dans la figure II.12), l'accumulation de temps thermique sur un pas de temps journalier permet d'enchaîner les différents stades phénologiques, pour lesquels différents processus physiologiques sont mobilisés. Les 4 stades clés sont : l'initiation florale (E2), le début de la floraison (F1), le début du remplissage des grains (M0) et la maturité physiologique (M3). Après l'élaboration de la surface foliaire, la biomasse est produite suivant l'équation de Monteih (1977) décrite ci-dessous, puis allouée aux grains avant de former le rendement.

$$Y = HI \int_{d=levée}^{récolte} RUE_d RIE_d PAR_d dt$$

où  $PAR_d$  correspond aux radiations photosynthétiquement actives incidentes chaque jour (MJ : Mega-joule),  $RIE_d$  à l'efficacité d'interception des radiations,  $RUE_d$  l'efficacité d'utilisation des radiations (g MJ<sup>-1</sup> : grammes de biomasse par Mega-joule),  $HI$  l'indice de récolte et  $Y$ , le rendement (g m<sup>-2</sup> : grammes par mètre carré).

Les variables d'entrée du modèle décrivent la météo, le sol, la conduite de culture et les caractéristiques variétales. Chaque variété dont on souhaite simuler le comportement est décrite par des paramètres spécifiques mesurés en serre ou champ, appelés paramètres « génotypiques », et qui sont en réalité des caractéristiques phénotypiques propres à chaque variété. Parmi ces paramètres, au nombre de 12, on trouve des variables de phénologie, d'architecture foliaire, d'allocation de la biomasse aux grains, et de réponse à la contrainte hydrique (Table II.6). Ils ont été choisis de manière à avoir suffisamment de variabilité entre géotypes ainsi qu'une sensibilité sur le modèle et être accessible à la mesure à des conditions de coût raisonnables. Un algorithme permet ensuite de simuler la croissance chaque jour du cycle. Ce modèle de croissance potentiel est modulé par deux principaux stress abiotiques : le stress hydrique et le stress azoté. Le stress hydrique affecte le rendement à travers 3 processus que sont l'expansion foliaire, la transpiration et l'accumulation de biomasse. Les sorties du

<b>Sol</b>	<i>profondeur</i>	<i>profondeur d'enracinement maximal (mm)</i>
	rh1	reliquats azotés
	rh2	reliquats azotés
	<i>Hcc</i>	<i>humidité massique à la capacité au champ (%)</i>
	<i>Hpf</i>	<i>humidité massique au point de flétrissement (%)</i>
	TC	taux de cailloux
da	densité apparente(g/cm3)	
Hini_C1	humidité massique au semis (%) horizon 0-30cm	
Hini_C2	humidité massique au semis (%) horizon 30-front racinaire	
<b>Conduite</b>	<i>densite</i>	<i>densité du peuplement (plantes/m2)</i>
	<i>jsemis</i>	<i>date semis</i>
	<i>jrecolte</i>	<i>date récolte</i>
	<i>apport_irri_i</i>	<i>irrigation (mm)</i>
	<i>date_irri_i</i>	<i>date de l'irrigation</i>
	<i>apport_ferti_i</i>	<i>apport de fertilisation (kg/ha d'azote minéral)</i>
	<i>date_ferti_i</i>	<i>date de fertilisation</i>
<b>Meteo</b>	<i>Tn</i>	<i>Température minimale de l'air (°C)</i>
	<i>Tx</i>	<i>Température maximale de l'air (°C)</i>
	<i>RG</i>	<i>Rayonnement global incident (MJ/m²)</i>
	<i>ETP</i>	<i>Evapotranspiration (Penman-Monteith) (mm)</i>
	<i>RR</i>	<i>Précipitations (mm)</i>
<b>Paramètre génétique</b>	variete	nom commercial
	HI	indice de récolte potentiel
	PHS	potentiel de photosynthèse relatif
	TLN	nombre de feuille potentiel
	LE	seuil de réponse de l'expansion foliaire à une contrainte hydrique
	TR	seuil de réponse de la conductance stomatique à une contrainte hydrique
	LLH	rang (depuis le sol) de la plus grande feuille du profil foliaire à la floraison :
	LLS	surface de la plus grande feuille du profil foliaire à la floraison (cm2)
	TTE1	somme de température (base 4.8) au stade bouton étoile depuis la levée
	TDF1	somme de température (base 4.8) à la floraison
	TDM0	somme de température (base 4.8) au debut de la maturité
	TDM3	somme de température (base 4.8) à la maturité physiologique
	K	coefficient d'extinction du rayonnement lors de la phase végétative (E1-F1)
OC	teneur en huile dans l'akène en conditions potentielles (%)	

**Table II.5 : Description des paramètres d'entrée du modèle (Ceux utilisés sont en couleur).**

environnement	année	profondeur estimée	hcc	hpf	densite	jsemis	jrecolte	irrigation	precipitation
AI08_I	2008	600			6.9	06/05/2008	23/09/2008	70	264
AI08_NI	2008	600			6.9	06/05/2008	16/09/2008	0	264
AI09_I	2009	600			6.9	24/04/2009	07/09/2009	80	137
AI09_NI	2009	600			6.9	24/04/2009	07/09/2009	0	137
CA10	2010	600	41.47	20.30	6.6	24/04/2010	11/09/2010	0	292
CHA10	2010	500	31.28	14.85	6.6	30/04/2010	02/10/2010	50	368.5
CO08_I	2008	400	22.15	10.23	6.5	22/05/2008	08/10/2008	98.5	138
CO08_NI	2008	400	22.75	10.90	6.5	22/05/2008	19/09/2008	0	138
CO09_I	2009	1000	20.52	9.60	6.5	19/05/2009	29/09/2009	102.96	134
CO09_NI	2009	1000	18.84	8.87	6.5	19/05/2009	29/09/2009	0	134
GA09_I	2009	1500	32.15	15.20	6.6	06/05/2009	09/09/2009	75	136
GA09_NI	2009	1500	32.15	15.20	6.6	06/05/2009	02/09/2009	0	136
LO10	2010	1500			6.9	15/04/2010	11/09/2010	0	168.6
SE10	2010	1000	37.10	18.10	6.5	29/04/2010	30/09/2010	0	97
VE09_I	2009	1500			6.9	07/05/2009	10/09/2009	80	119.67
VE09_NI	2009	1500			6.9	07/05/2009	10/09/2009	0	119.67
VE10	2010	1500			6.9	07/05/2010	10/09/2010	35	204

**Table II.6: Valeurs des variables de conduite de culture et de description du sol pour les environnements modélisés**

modèles consistent en différentes variables de performance, telles que le rendement et la teneur en huile, mais aussi par exemple certains indicateurs de stress subi par la culture. La version utilisée dans le cadre de cette thèse est celle implémentée dans le logiciel ModelMaker. Récemment, le modèle a été retranscrit sur la plateforme RECORD (REnovation et COoRDination de la modélisation de cultures pour la gestion des agro-écosystèmes: <http://www.inra.fr/record>).

### II.3.3 Caractérisation des environnements avec le modèle SUNFLO

#### II.3.3.1 Matériels et méthodes

Pour caractériser le stress hydrique dans chaque environnement, un ensemble de données climatiques ainsi que de données liées au sol et à la conduite de culture a été recueilli. La table II.5 présente la définition des paramètres utilisés en entrée pour les simulations dans le modèle SUNFLO. Les variables disponibles sont mises en évidence par de la couleur.

Concernant le sol, plusieurs estimations de la profondeur ont été fournies par les agriculteurs et expérimentateurs. La différence entre l'humidité à la capacité au champ, (sol ressuyé dans des conditions où le drainage est assuré librement) et l'humidité au point de flétrissement (l'eau est retenue avec une intensité supérieure aux forces de succion des racines) permet de calculer la réserve utile en prenant en compte la densité. Ces données d'humidité ont été obtenues grâce à des analyses de sol, sur seulement 8 environnements (Table II.6).

Concernant les données climatiques, elles ont été recueillies à partir de pluviomètres et sondes de températures disponibles sur certains essais et/ou à partir de la base de données climatiques gérée par l'unité de service Agroclim de l'INRA (<http://www6.paca.inra.fr/agroclim>). L'origine des données climatiques est détaillée dans la table II.7. Même si la plupart des données de température et de précipitations ont été recueillies sur place, les variables d'ETP (Evapotranspiration Potentielle) et de rayonnement global ont été récupérées pour la plupart sur des stations plus éloignées (jusqu'à 38 km).

Trois variétés décrites pour un ensemble de paramètres « génotypiques » (Table II.8) dans le modèle SUNFLO figurent parmi les quatre témoins des expérimentations. Les simulations ont été lancées pour chacune de ces variétés indépendamment, dans chaque environnement. Au total, 51 simulations ont été effectuées (trois variétés/17 environnements) avec ModelMaker en utilisant la version SUNFLO\_v1.

<b>Origine des données:</b>				
<b>Environnement</b>	<b>Températures</b>	<b>ETP</b>	<b>Précipitations</b>	<b>Rayonnement global</b>
<b>CO08_I</b>	Station (9 km)	Station (26 km)	Pluviomètre (0 km)	Station (26 km)
<b>CO08_NI</b>				
<b>CO09_I</b>	Station (9 km)	Station (26 km)	Pluviomètre (0 km)	Station (26 km)
<b>CO09_NI</b>				
<b>AI08_I</b>	Station (0km)	Station (28 km)	Station (0km)	Station (28 km)
<b>AI08_NI</b>				
<b>GA09_I</b>	Station ( 3km)	Station (38km)	Station (3km)	Station (38 km)
<b>GA09_NI</b>				
<b>AI09_I</b>	Station (0km)	Station (28 km)	Station (0km)	Station (28 km)
<b>AI09_NI</b>				
<b>VE09_I</b>	Station (0km)	Station (30km)	Station (0km)	Station (30 km)
<b>VE09_NI</b>				
<b>CHA10</b>	Station (6km)	Station (6km)	Station (6km)	Station (6km)
<b>CA10</b>	Station (3km)	Station (17km)	Station (3km)	Station (17km)
<b>LO10</b>	Station (0 km)	Station (0 km)	Station (0 km)	Station (0 km)
<b>VE10</b>	Station (11 km)	Station (11 km)	Station (11 km)	Station (11 km)
<b>SE10</b>	Station (0km)	Station (20km)	Pluviomètre (0 km)	Station (38km)

**Table II.7 : Origine des données météo**

<b>Paramètre</b>	<b>résumé</b>	<b>MELODY</b>	<b>PEGASOL</b>	<b>PACIFIC</b>
HI	Indice de récolte	0.42	0.44	0.39
PHS	Photosynthèse	1	1	1
TLN	Nb de feuilles	28.7	25.3	23.5
LE	expansion foliaire	-3.896	-3.687	-3.359
TR	conductance stomatique	-10.699	-9.998	-10.124
LLH	rang de la plus grande feuille	17.4	25.3	17.5
LLS	surface de la plus grande feuille	537	17.4	420
TTE1	Somme t°C levée-bouton étoilé	542	522	531
TDF1	Somme t°C floraison	941	906	922
TDM0	Somme t°C M0	1188	1153	1169
TDM3	Somme t°C M3	1751	1721	1722
K	coefficient d'extinction	0.838	0.856	0.847
OC	teneur en huile	45.6	47.3	46.5

**Table II.8 : Valeurs des paramètres « génétiques » des 3 variétés témoins**

Pour chaque simulation, le rendement simulé a été extrait et comparé au rendement observé moyen pour chaque variété dans chaque environnement. Pour mesurer la précision de prédiction, les carrés moyens de l'erreur (RMSE : root mean square error) ont été utilisés.

$$RMSE = (\sqrt{E(\hat{\mu} - \mu)^2})$$

Il s'agit de la racine carrée de l'espérance du carré de la différence entre la valeur prédite ( $\hat{\mu}$ ) par le modèle SUNFLO et la valeur observée ( $\mu$ ). Cette RMSE a été estimée par son estimateur empirique sur l'ensemble des rendements observés et prédits. Le RRMSE (RMSE relatif), ratio du RMSE par l'espérance de la valeur observée, a également été calculé.

Afin de minimiser la valeur du RMSE, la profondeur du sol a été ajustée sur certains environnements, tout en restant dans des valeurs acceptables. Cette variable est en général difficile à observer au champ ; il semble qu'elle a un impact essentiel sur le rendement simulé dans le modèle.

Plusieurs indicateurs de stress ont ensuite été extraits du modèle pour chaque combinaison variété - environnement. Parmi ces indicateurs, le rapport ETR/ETM permet d'avoir une bonne évaluation du stress hydrique. L'ETR, l'évapotranspiration réelle, est la quantité d'eau évapotranspirée par une culture et dépend de l'état hydrique du sol et du stade phénologique. L'ETM est l'évapotranspiration maximale, c'est-à-dire la quantité maximum d'eau évaporée par le couvert placé dans de bonnes conditions hydriques.

$$\frac{ETR}{ETM} = (Evj + vTR)/(Evj + (1.2 * ETP * Ei))$$

où  $Evj$  est la vitesse d'évaporation (mm/jour),  $vTR$  est la vitesse de transpiration,  $ETP$  est l'évapotranspiration potentielle de Penman Monteith et  $Ei$  est l'efficacité d'interception du couvert. Lorsque ce rapport est inférieur à 0.6, un jour de stress est cumulé en sortie du modèle, ce qui permet d'obtenir un nombre de jours de stress (JS) sur le cycle complet et pour 3 phases du cycle : JS1, au stade végétatif (entre les stades A2 et F1), JS2, autour de la floraison (entre les stades F1 et M0) et JS3 au moment du remplissage du grain (entre les stades M0 et M3).

En plus de ces indicateurs liés à la contrainte hydrique, la contrainte thermique a également été estimée pour chaque combinaison variété - environnement et à plusieurs stades du cycle.





Dans le modèle SUNFLO, la température a un effet direct (en plus de son effet sur les degrés jours) sur l'efficacité de l'utilisation des radiations (RUE, g MJ<sup>-1</sup>) et sur la minéralisation de l'azote. La RUE est ainsi affectée par un facteur thermique FT calculée de la manière suivante (Casadebaig *et al.*, 2010) :

$$Tm \left( \frac{1}{Tol - Tb} \right) - \left( \frac{Tb}{Tol - Tb} \right) \text{ si } Tm < Tol$$

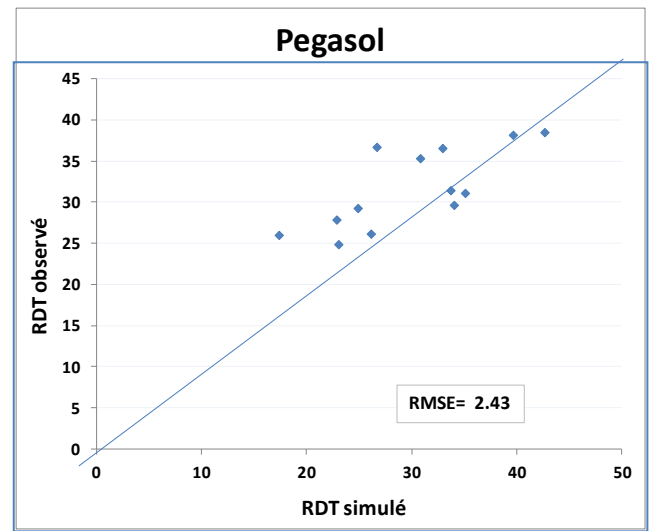
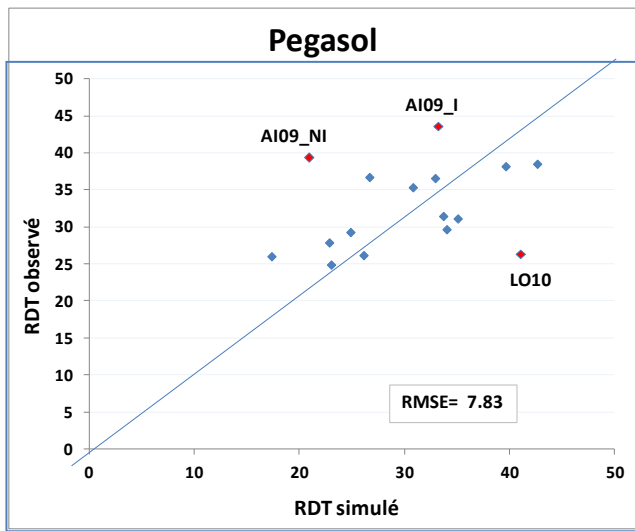
$$Tm \left( \frac{1}{Tol - Tc} \right) - \left( \frac{Tb}{Tol - Tb} \right) \text{ si } Tm > Tou$$

$$1 \text{ si } Tol < Tm < Tb$$

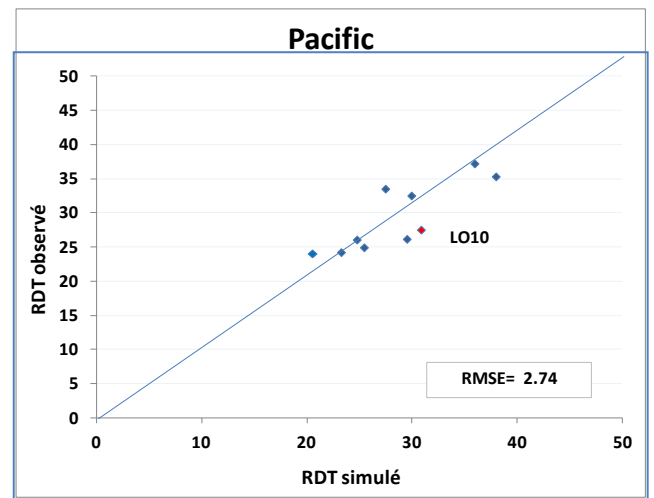
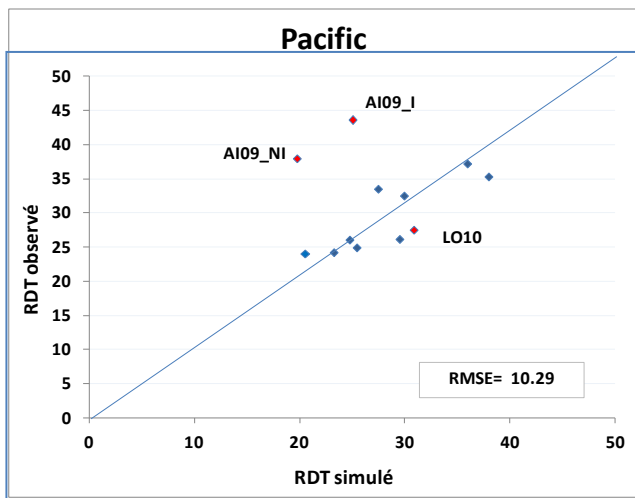
où  $Tm$  est la température moyenne,  $Tb$  est la température de base (4.8°C),  $Tol$  est la température optimale basse ( $Tol=20^\circ\text{C}$ ),  $Tou$  est la température optimale haute ( $Tou=28^\circ\text{C}$ ) et  $Tc$  est la température critique ( $Tc=37^\circ\text{C}$ ).

L'équation précédente permet donc de calculer une valeur de FT par jour, qui, lorsqu'elle est optimale, c'est-à-dire lorsque la température est entre 20°C et 28°C, n'a pas d'impact négatif sur la RUE et est égale à 1. La somme de l'inverse des valeurs de FT le long du cycle ou sur les 3 phases de développement (A2-F1, F1-M0 et M0-M3) a permis d'obtenir les indicateurs IFT, IFT1, IFT2 et IFT3 respectivement. Ces indicateurs peuvent traduire à la fois un stress dû à des températures élevées, supérieures au seuil critique  $Tc$  ou un stress froid lorsque les températures sont inférieure à  $Tol$ .

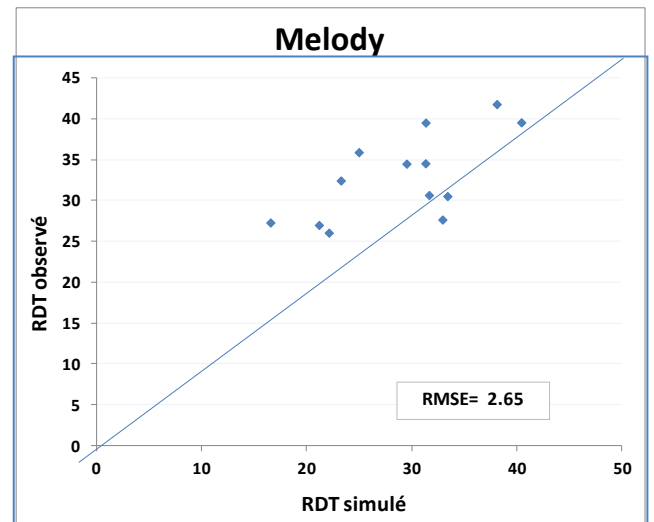
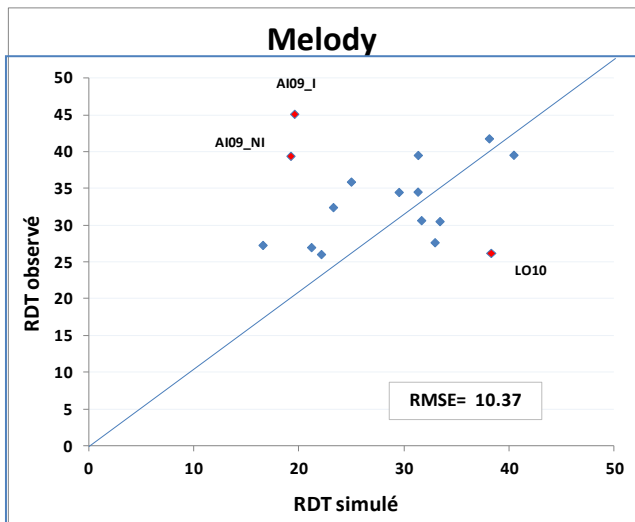
L'impact du stress hydrique sur le rendement dépend de la phase du cycle. En tournesol, il a été montré que cet impact était de l'ordre de 10 à 25 % de perte si le stress a lieu pendant la phase végétative et jusqu'à 50% autour de floraison (Rauf *et al.*, 2008). Les études concernant l'impact du stress thermique chez le tournesol sont plus rares. Rondanini *et al.*, (2003) ont montré une baisse de 40% du rendement en grains après exposition des capitules à une température supérieure à 40°C pendant 7 jours consécutifs au moment du remplissage des grains (M0M3). Le pourcentage d'huile a été réduit de 30%. Quant au stress froid, il partage certaines conséquences des stress hydrique et thermique chaud à travers les dommages membranaires qu'il peut causer (cf Chapitre I). Chez le tournesol, il est d'importance non négligeable notamment au moment de la germination et des premiers stades d'émergence ; pourtant il n'a été que très peu étudié (Skoric *et al.*, 2009). Le stress consécutif au froid n'est



a



b



c

Figure II.13 : Rendement observé en fonction du rendement simulé avant (à gauche) et après (à droite) ajustement de la profondeur du sol pour les 3 variétés : PEGASOL(a), PACIFIC(b) et MELODY(c).

pas l'objet de ce travail mais l'index FT nous permettra de savoir si les environnements ont subis un stress thermique et de quelle nature il est.

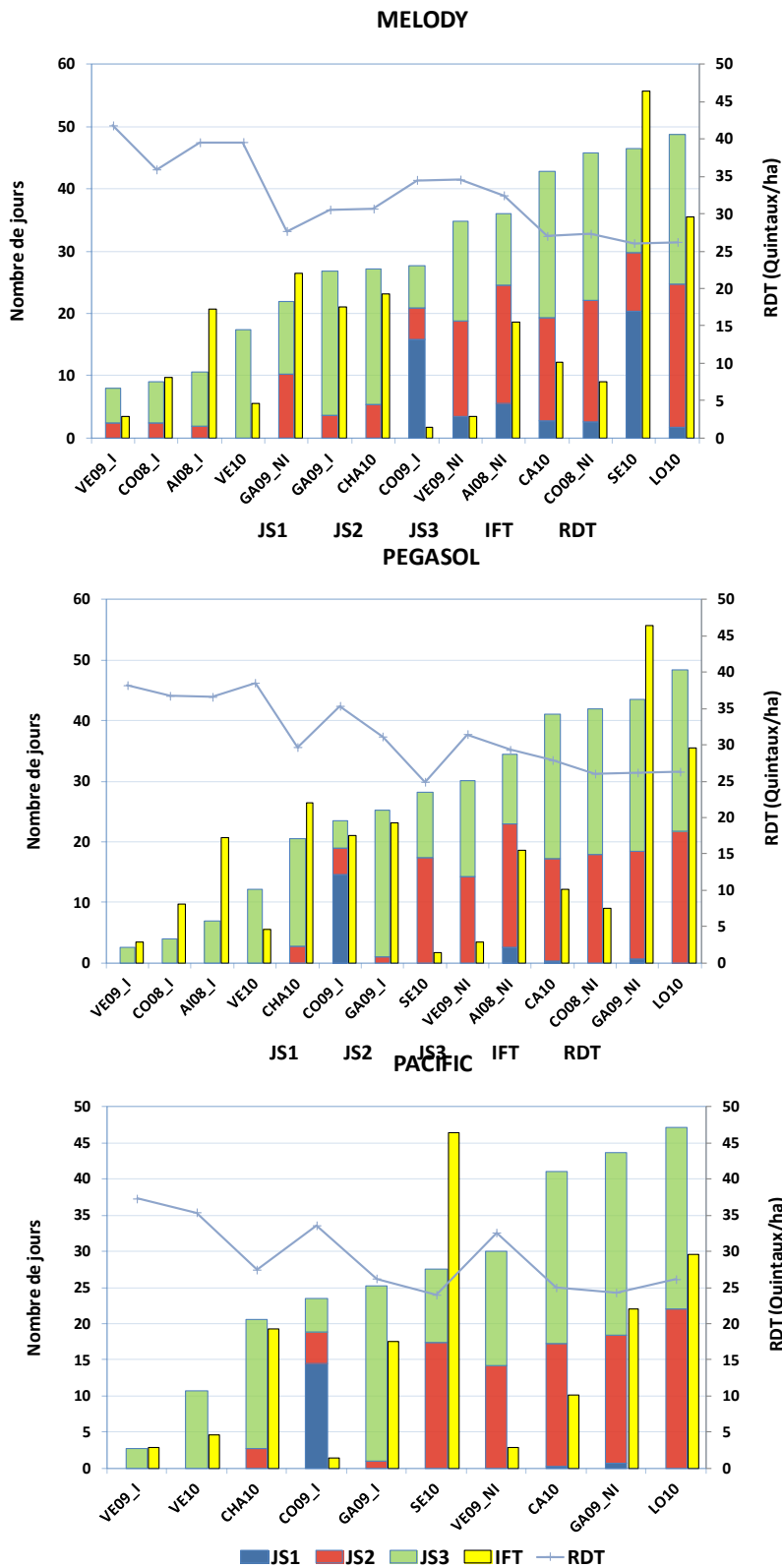
Afin de mesurer l'impact des indicateurs de stress hydrique et thermiques calculés à différents stades du cycle sur plusieurs variables liées à la productivité (RDT, RDTH, PMG et Nbgrains), des analyses en composantes principales ont été effectuées pour chaque variété sur les indicateurs de stress calculés dans chaque environnement. Des régressions multiples ont ensuite été menées en utilisant la fonction step de R qui permet de sélectionner des variables grâce au critère AIC. Trois environnements sur 17 ont été exclus des analyses : AI09\_I et AI09\_NI car mal prédits par le modèle SUNFLO et CO09\_NI dont l'effet génotype n'était pas significatif pour le rendement.

### II.3.3.2 Résultats et discussion

#### - Qualité de prédiction du modèle SUNFLO

Les graphes des rendements observés en fonction des rendements simulés sont présentés figure II.13 pour chaque variété. La variété PACIFIC n'ayant pas été évaluée sur les 4 environnements de 2008 (témoin INEDI à la place) présente moins de points de comparaison observé/simulé. Les RMSE sont très similaires entre les 3 variétés. Ils varient de 7.83 à 10.37 en considérant les simulations à partir des profondeurs de sol estimées par les sélectionneurs ou agriculteurs. Les RRMSE s'étendent de 24 à 33%, ce qui est légèrement supérieur aux valeurs calculées dans Casedebaig *et al.*(2010) (4 à 30%).

Un des lieux se distingue par son écart entre rendement observé et rendement prédit, avec jusqu'à 25 qtx/ha de plus pour le rendement observé. Les deux environnements correspondant à ce lieu (AI09\_I et AI09\_NI), identifiés en rouge sur les graphes, ont été enlevés des analyses car même en augmentant au maximum la profondeur de sol dans le modèle, il n'y a quasiment pas d'impact sur le rendement simulé. Il semble qu'il manque donc certains paramètres pour caractériser ces deux environnements à très fort potentiel. LO10 apparaît également mal prédit par le modèle, avec à l'inverse un rendement simulé plus fort que le rendement observé. La profondeur de sol a été diminuée dans cet environnement de 1500 à 700 mm. Au total, 8 environnements ont nécessité des ajustements sur la profondeur avec un écart maximum de 80 cm et 6 environnements n'ont pas été ajustés. Après ajustements les RMSE varient de 2.43 à 2.74 et les RRMSE sont inférieurs à 9%.



**Figure II.14** Distribution des indicateurs de stress hydrique selon le stade phénologique (JS1: avant floraison), JS2 autour de la floraison et JS3 après la floraison) ainsi que le cumul tout au long du cycle de stress thermique (IFT). Le rendement observé sur chacun des génotypes est également représenté.

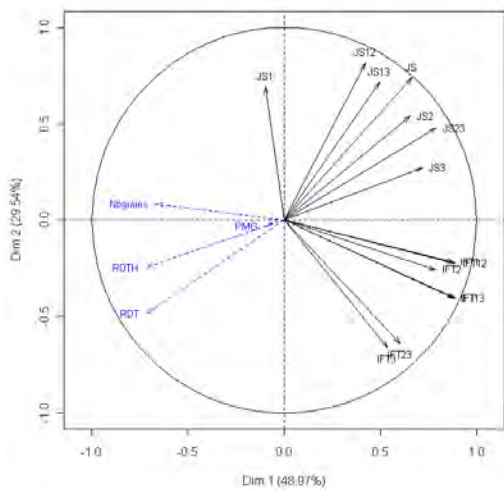
## - Indicateurs de stress

Après avoir minimisé les RMSE, plusieurs indicateurs de stress ont été extraits du modèle. La distribution du nombre de jours de stress hydrique avant (JS1), autour (JS2) et après (JS3) floraison pour l'ensemble des environnements est représentée figure II.14 pour chacune des 3 variétés. Le stress avant floraison (JS1) est très peu présent sur ce réseau expérimental à part sur un environnement : CO09\_I. Cet environnement est un des rares pour lequel des symptômes de flétrissement avaient été observés. Le stress post floraison est lui présent sur la plupart des environnements et varie de 2 à 26 jours. Le stress autour de la floraison s'étend jusqu'à un maximum de 22 jours et semble différencier 2 groupes d'environnements, le premier groupe, rassemblant les lieux irrigués et le second groupe, les lieux non-irrigués. L'irrigation a donc bien permis d'atténuer le stress vers la floraison. D'une variété à l'autre le classement des lieux en prenant en compte le cumul de stress varie légèrement. Cependant, le classement des 4 environnements les moins stressants est stable. Les différences de classement entre MELODY et PEGASOL proviennent essentiellement de GA09\_I et GA09\_NI qui sont ressentis comme plus stressants par la variété PEGASOL, ainsi que SE10 qui lui est senti comme moins stressant. Il y a donc une interaction variété - environnement probablement plus forte pour ces environnements. Quant à PACIFIC, sur les lieux 2009 et 2010, son classement est identique à celui de PEGASOL.

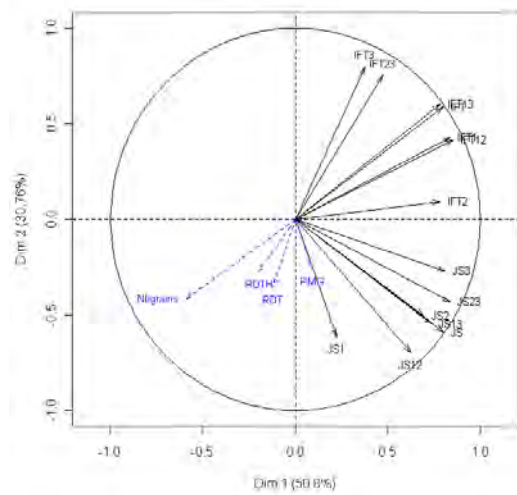
La figure II.14.b représente aussi la distribution de l'indicateur de stress thermique (IFT) pour la variété MELODY. Le profil de stress thermique est très similaire entre les variétés. En effet, cet indicateur dépend essentiellement des températures et le facteur variétal n'est dû qu'à la durée des stades.

Le stress thermique avant floraison est le plus fort et varie de 0.6 à 23 unités selon les environnements, le stress autour de la floraison est quasiment absent et en post floraison, il est présent sur les environnements ayant déjà accumulé le plus de stress thermique en pré floraison. Ces environnements sont donc caractérisés par des températures non optimales qui commencent à être critiques pour les processus de photosynthèse et donc la RUE. Lorsque l'on regarde quel type de stress thermique est limitant, il s'agit principalement du stress froid avec des températures inférieures à 20°C en moyenne quasiment sur tout le cycle à part autour de la floraison.

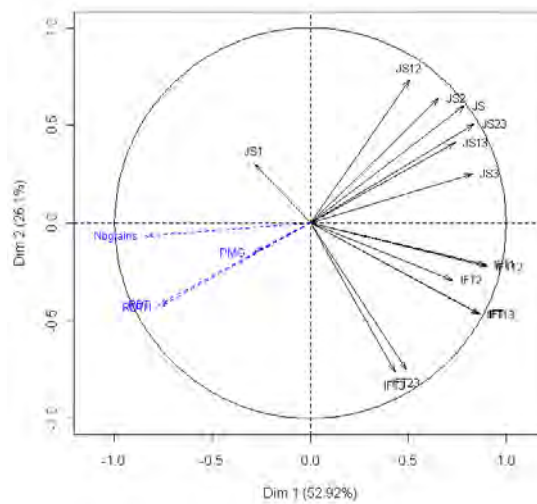
La figure II.14. présente la comparaison des indicateurs de stress hydrique et thermique cumulés le long du cycle en relation avec le rendement pour chacune des variétés. Pour un



a



b



c

**Figure II.15: Cercle des corrélations de l'ACP des indicateurs de stress hydriques et thermiques et des variables de réponse (RDT : rendement en grains observé, RDTH : rendement en huile observé, Nbrgrains : nombre de grains observé, PMG : poids de mille grain observé). Les % de variances expliquées par les premières et deuxièmes composantes de l'ACP sont indiqués en abscisse et en ordonnée. Les génotypes sont : a. MELODY b. PACIFIC. c. PEGASOL Tous les indicateurs de stress possibles sont représentés : JS, pour le stress hydrique cumulé tout au long du cycle ou sur les stades avant et autour floraison (JS12), autour et après floraison (JS23), avant et après floraison (JS13) ou non cumulés (JS1 ; JS2 et JS3). La même nomenclature s'applique aux indicateurs de stress thermique (IFT, IFT1, IFT2, IFT12, IFT13, IFT23).**

stress hydrique croissant, le rendement diminue. Par contre, le stress thermique seul ne produit pas la même tendance, étant aussi marqué sur des lieux ayant un fort nombre de jours de stress hydrique ou non. Cependant, on peut noter que LO10 cumule à la fois une forte valeur de stress hydrique et de stress thermique. Ces premières observations laissent penser que le stress thermique seul ne permet pas d'expliquer la réduction du rendement alors que le stress hydrique semble avoir une meilleure capacité de prédiction.

- Choix des indicateurs de stress pour prédire la productivité

Les ACP présentées figure II.15 ont été réalisées à partir des indicateurs de stress hydrique et thermique cumulés ou non sur les différents stades phénologiques (levée-F1, F1M0 et M0M3). Les variables de réponse (rendement en grains et en huile, nombre de grains et PMG) ont été positionnées dans un second temps et ne participent donc pas à l'élaboration des composantes principales. Quelle que soit la variété, on remarque que les indicateurs se regroupent par type de stress: thermique ou hydrique. Ces deux types de stress sont peu corrélés et caractérisent donc différemment les environnements, le stress hydrique étant probablement indépendant d'un stress thermique froid. A l'intérieur de chacun des groupes, les variables sont très corrélées, le stade phénologique ne semble donc pas apporter d'influence majeure. Cependant, le stress hydrique avant floraison (JS1) se différencie des autres variables, surtout pour la variété PEGASOL où il est décorrélé des autres indicateurs de stress hydrique et des variables phénotypiques. Pour PACIFIC, il est corrélé au PMG mais cet indicateur n'est pas bien représenté dans ce plan factoriel. En effet le stress avant floraison est très peu présent dans le dispositif et a donc peu de poids. Plusieurs variables de réponses sont également mal prédites dans ces plans factoriels. C'est le cas du PMG pour les 3 variétés. Les processus aboutissant à ce caractère sont peut-être moins influencés par le stress hydrique ou thermique tel qu'il est caractérisé ici. Les variables de rendement apparaissent également moins bien prédites pour PACIFIC. Pour MELODY et PEGASOL, le rendement est corrélé négativement aux indicateurs de stress hydrique. Quant au nombre de grains, il est corrélé négativement aux indicateurs de stress thermique pour MELODY et PACIFIC alors qu'il reste inexpliqué par les indicateurs pour PEGASOL.

Afin de sélectionner le meilleur indicateur, ou combinaison d'indicateurs, pour prédire le rendement (huile et grains) et le nombre de grains, nous avons effectué des régressions multiples de ces variables de réponse sur deux variables explicatives : le nombre de jours de stress hydrique cumulé sur l'ensemble du cycle (JS) et le nombre de jours de stress thermique cumulé sur l'ensemble du cycle (IFT) tel que cela était suggéré par l'ACP. L'utilisation de

Variable de réponse	Génotype	Nombre d'environnements	Indicateurs	P-value	Coefficient	R <sup>2</sup> ajusté
RDT	MELODY	14	JS	4.99E-04	-0.29	0.70
			IFT	0.13	-0.17	
	PEGASOL	14	JS	3.56E-05	-0.28	0.75
	PACIFIC	10	JS	0.07	-0.20	0.58
IFT			0.24	-0.19		
RDTH	MELODY	13	JS	0.03	-0.10	0.56
			IFT	0.07	-0.14	
	PEGASOL	14	JS	5.18E-05	-0.16	0.74
	PACIFIC	10	JS	0.02	-0.13	0.69
IFT			0.20	-0.10		
Nbgrains	MELODY	8	IFT	0.01	-1268.8	0.56
	PEGASOL	9	IFT	0.0989	- 821	0.66
			JS	-394.8	0.1340	
	PACIFIC	6	IFT	-973.4	0.12680	0.84
JS			-337.3	0.32908		

**Table II.9 : Résultats des régressions multiples des variables de réponse sur les indicateurs de stress JS (stress hydrique) et IFT (stress thermique).** Pour chaque combinaison variable-génotype, les covariables du meilleur modèle sont spécifiées avec leur p-value et le coefficient de régression correspondant. La variance expliquée par le modèle (R<sup>2</sup> ajusté) est également précisé.



toutes les variables (à différents stades) risquait d'entraîner un sur-apprentissage étant donné le faible nombre d'environnements (de 9 à 14).

La table II.9 présente pour chaque variable expliquée le meilleur modèle explicatif, basé sur l'AIC. Par exemple, pour chaque unité de stress hydrique supplémentaire, lorsque l'IFT est fixé, le rendement diminue de 0.29 quintaux pour la variété MELODY. D'après cette table, le modèle comprenant les 2 covariables JS et IFT est le plus souvent le meilleur, quelle que soit la variété et le caractère. La variation phénotypique expliquée par les modèles s'étend de 0.56 à 0.84, ce qui témoigne d'une bonne qualité de prédiction. Pour le rendement en grain et en huile, la variable JS est nécessaire voire suffisante (PEGASOL) alors que pour le nombre de grains, qui se présentait de manière bien différenciée sur l'ACP, c'est plutôt le facteur thermique qui est nécessaire au modèle, voire suffisant pour MELODY.

MELODY est connue pour avoir une stratégie de réponse au stress hydrique différente de PEGASOL et PACIFIC. Ces 2 dernières variétés ont tendance à continuer de produire (stratégie « deshydratation tolérance ») malgré la présence de stress hydrique, tandis que MELODY a une stratégie plus conservative (« deshydratation avoidance »), avec une fermeture des stomates plus rapide pour maintenir le statut hydrique.

Cette différence de stratégie ne ressort pas dans nos résultats. Seul le nombre de grains, où le rendement de MELODY n'est expliqué que par le facteur thermique et non par les deux types de stress, s'illustre différemment. Cette variété pourrait être plus sensible au froid que les autres.

Cette analyse nous permet donc d'identifier deux covariables environnementales qui seront utiles pour prédire la productivité des lignées du panel et comparer leur réponse à un stress croissant. On peut noter que les dates de floraison sont très proches entre variétés témoins (de 194 pour PACIFIC à 198 pour MELODY), ce qui signifie que le stress a été appliqué au même stade entre génotypes.

#### *II.3.4* Mise en place d'un index synthétique de réponse au stress pour le panel d'association

##### *II.3.4.1* Matériels et méthodes

Suite aux simulations avec le modèle SUNFLO, nous avons donc identifié deux indicateurs de stress qui permettent de prédire le rendement grain, le rendement en huile et le nombre de



grains des variétés témoins. Ces indicateurs sont utiles à la description des environnements et à leur classification (Figure II.14) mais ils permettent aussi de comparer la réponse des génotypes à un stress croissant grâce à l'exploitation des pentes de la régression des variables de productivité sur ces indicateurs. Les pentes fortes traduisent une plus forte sensibilité au stress.

Les BLUP des variables phénotypiques sélectionnées (RDT, RDTH, Nbgrains) sur l'ensemble des environnements ont été estimés pour les 384 lignées du panel (cf. section II.2). Les individus ayant plus de 50% de données manquantes (huit environnements) ont été éliminés. Le reste des données manquantes a été inférés pour chaque caractère en utilisant le package missMDA (<http://cran.r-project.org/web/packages/missMDA/>) qui utilise l'algorithme itératif de l'ACP décrit dans Kiers (1997).

Le modèle de régression suivant a été appliqué pour chaque génotype du panel en utilisant la fonction lm de R.

$$Y_{ij} = a_i + b_i JS_j + c_i IFT_j + \varepsilon_{ij}$$

avec  $Y_{ij}$  le BLUP de la variable de réponse pour le génotype  $i$ ,  $a_i$ , la variable de réponse potentielle,  $JS_j$ , la moyenne du nombre de jours de stress hydrique des trois variétés témoins pour l'environnement  $j$ ,  $b_i$ , la pente associée au stress hydrique,  $IFT_j$ , la moyenne du stress thermique des trois variétés témoins pour l'environnement  $j$ ,  $c_i$ , la pente associée au stress thermique,  $\varepsilon_{ij}$ , l'erreur résiduelle.

Les valeurs de  $R^2$  ajusté ont été extraites de chaque modèle, ainsi que les pentes et ordonnées à l'origine. Une analyse de covariance a été menée afin de tester l'effet de la paire de testeurs utilisée (83HR4gms/FS71501 sur 8 environnements et SOLR0001M/AT0521 sur 9 environnements) sur les pentes des régressions selon le modèle suivant :

$$Y_{ij} = a_i + b_i JS_j * Groupe_k + \varepsilon_{ij}$$

Les covariables sont identiques au modèle précédent sauf qu'il n'y a que le stress hydrique comme variable explicative et  $Groupe_k$  correspond au groupe d'environnements ayant la même paire de testeurs (groupes 1 et 2).



### II.3.4.2 Résultats et discussion

Quatorze environnements (au maximum) sur les 17 disponibles ont été utilisés : AI09\_I et AI09\_NI ont été éliminés car mal prédits par le modèle SUNFLO et CO09\_NI a été éliminé car l'effet génotype pour le rendement n'y était pas significatif.

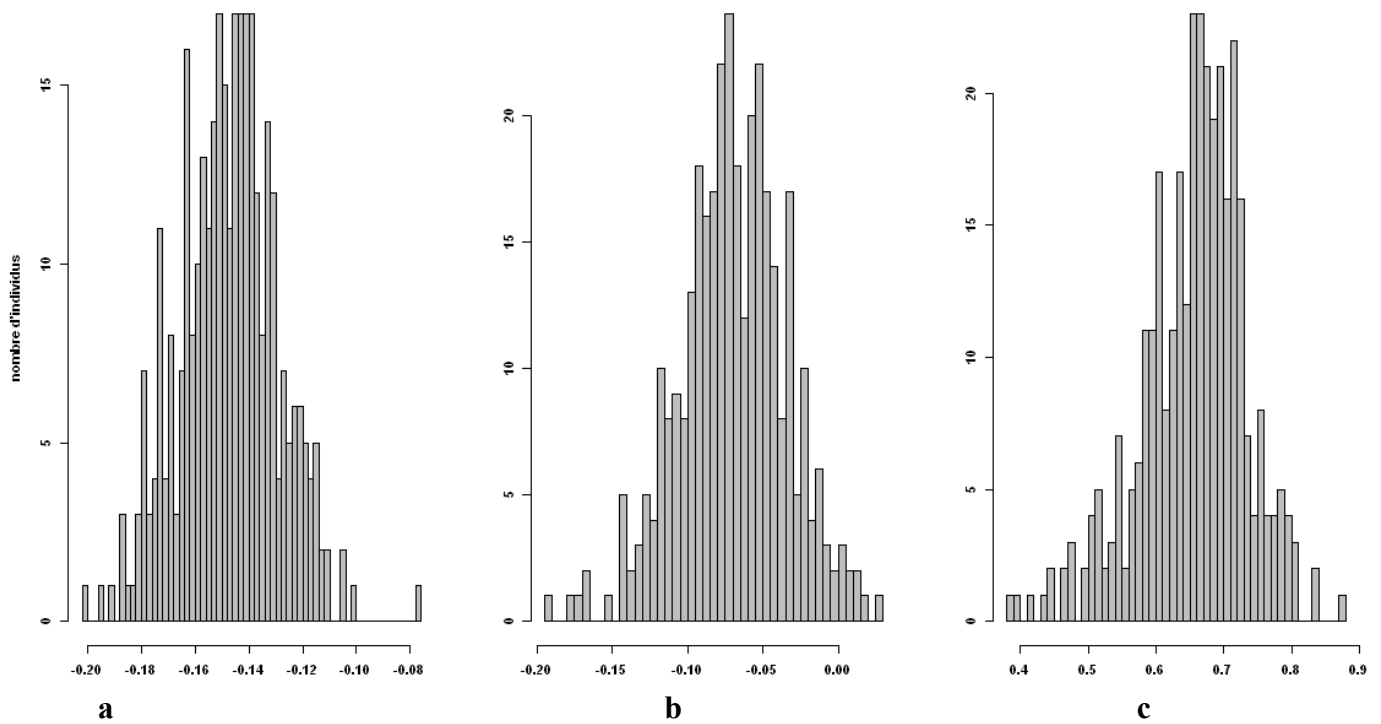
#### - Rendement en grain

Au total, pour le rendement, 337 hybrides (testeur - lignées du panel) présentant moins de sept environnements avec des données manquantes ont été conservés pour les analyses.

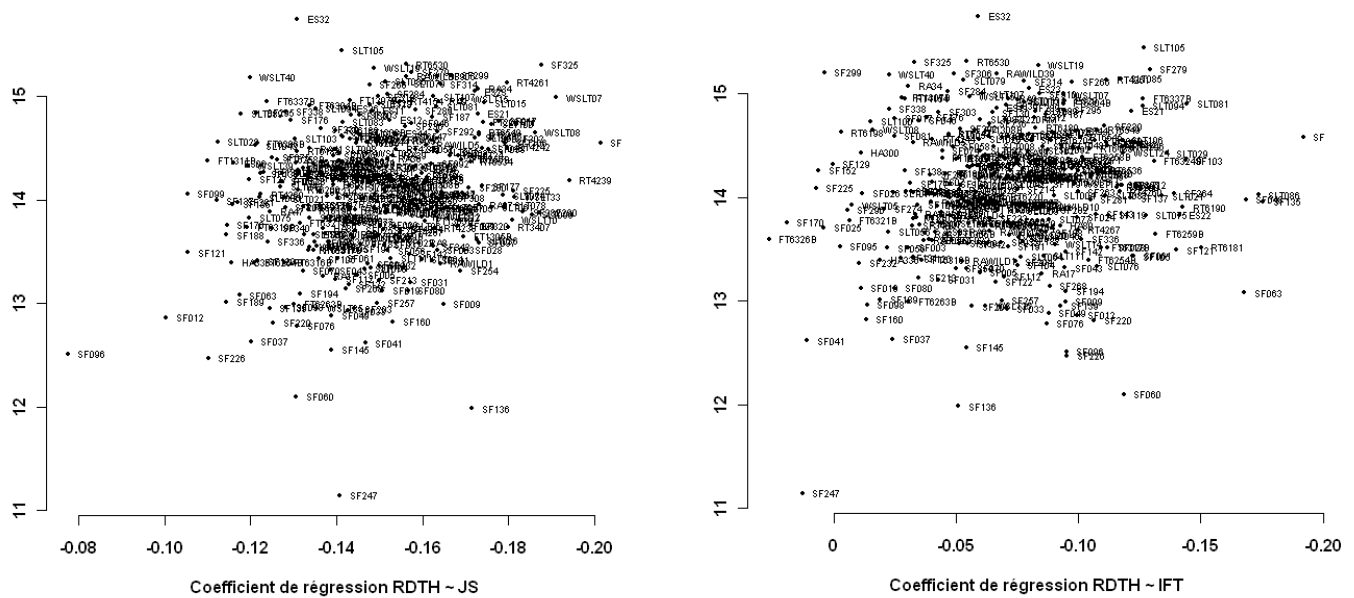
Même si le modèle IFT + JS a été sélectionné comme le plus pertinent pour prédire le rendement dans le paragraphe précédent, les  $R^2$  ajustés sont légèrement plus faibles que ceux des modèles ne comprenant que la covariable JS. Quant aux régressions avec l'IFT comme seule covariable, leur  $R^2$  moyen est très faible (de l'ordre de 0.08). Nous ne présenterons donc ici que les résultats du modèle avec une seule covariable : JS

Les coefficients de régression du rendement en grain (RDT) en fonction de JS varient de -0.40 à -0.19 selon les lignées du panel (coefficient moyen sur l'ensemble des lignées = -0.29). Globalement, la qualité des régressions est bonne. Le  $R^2$  ajusté moyen s'élève à 0.57 et varie de 0.31 à 0.80 (Figure II.16) et toutes les p-values associées à la covariable JS sont significatives.

Dans ce modèle de régression, l'ordonnée à l'origine traduit le rendement optimal au niveau de stress hydrique minimum qui s'est exercé dans ce réseau d'essai. Ce rendement optimal varie ici entre 30 et 40 quintaux/ha environ. Il existe une forte corrélation entre le rendement optimal et le coefficient de régression (corrélation de Pearson de -0.74). Plus le rendement optimal est élevé, plus le coefficient de régression est faible. Ceci est dû à la corrélation statistique des estimateurs. En centrant les covariables explicatives, cette corrélation statistique est réduite à zéro. Ainsi, après avoir centré la covariable JS, la corrélation pente/ordonnée à l'origine n'est plus que de -0.27 mais reste significative (p-value =  $3 \cdot 10^{-07}$ ) (Figure II.17). Elle peut traduire un phénomène biologique réel. Les génotypes à rendement potentiel élevé sont également les plus sensibles lorsque les conditions deviennent très stressantes (la vigueur augmente la demande évaporative). En centrant les covariables, les coefficients de régression restent inchangés mais l'ordonnée à l'origine est modifiée et doit s'interpréter différemment. En effet ce n'est plus la valeur de la variable dépendante quand les variables explicatives sont à zéro mais quand



**Figure II.18 : Distribution des coefficients de régressions pour JS (a) et IFT (b) et des R2 ajustés (c) selon les lignées, pour le modèle expliquant le rendement en huile en fonction des indicateurs JS et IFT.**



**Figure II.19 : Relation entre l'ordonnée à l'origine et les coefficients de régression (JS et IFT de gauche à droite) dans le modèle centré par la moyenne du stress (RDTH).**

celles-ci sont égales à la moyenne du stress. Ce qui explique que dans notre cas, la variance diminue et n'est plus que d'environ sept quintaux au lieu de dix ; tout le potentiel génétique ne s'exprime pas autant lorsqu'un stress plus important est présent.

Puisqu'il y a une corrélation biologique entre le rendement potentiel moyen et la pente de régression il est intéressant de chercher à sélectionner des individus maximisant le rendement potentiel tout en ayant un coefficient de régression le plus faible possible. Cette stratégie s'illustre bien sur le graphe présenté figure II.17. L'axe de variation représenté par la flèche traduit la différence entre les génotypes pour cette combinaison de potentialité de rendement et de stabilité. Il est donc intéressant d'exploiter l'index de sélection qui résulte de cet axe en génétique d'association. Pour l'obtenir, nous avons extrait la deuxième composante principale de l'ACP réalisée sur 2 colonnes : le rendement potentiel et le coefficient de régression.

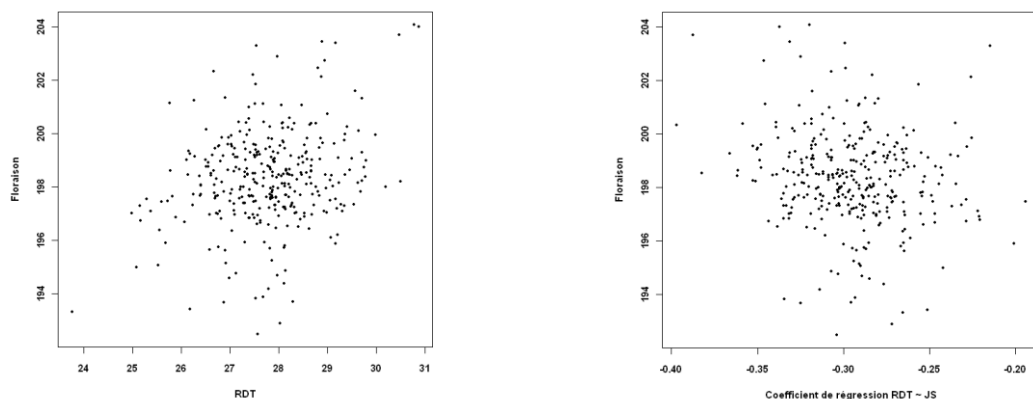
#### - Rendement en huile

Après le tri des données manquantes, 337 hybrides (testeur - lignées du panel) ont été sélectionnées pour les régressions. Contrairement au rendement en grains, le modèle JS + IFT apporte une meilleure prédiction que le modèle JS. Le  $R^2$  ajusté moyen pour ce modèle est de 0.67, le rendement en huile est donc mieux expliqué par les indicateurs de stress sélectionnés que le rendement grain. Le coefficient de régression pour JS varie de -0.20 à -0.07 et de -0.19 à 0.03 pour IFT (Figure II.18). Après avoir centré les covariables, le rendement en huile à niveau de stress moyen varie entre 11 et 16 quintaux/ha. De même les corrélations entre JS ou IFT et l'ordonnée à l'origine passent de, respectivement -0.57 et -0.39 à -0.23 et -0.11, mais elles restent significatives. Les relations entre le rendement en huile potentiel et chacune des covariables est présenté figure II.19.

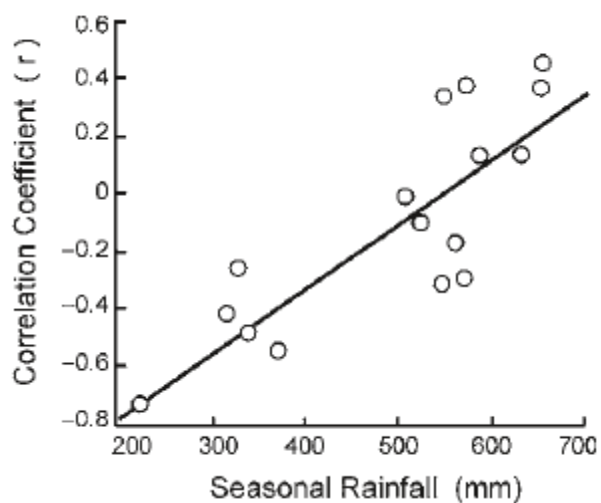
De la même façon que pour le rendement, les coordonnées des lignées sur l'axe de variation perpendiculaire à cette corrélation sera exploité comme variable synthétique supplémentaire en génétique d'association.

#### - Nombre de grains

Au total, 352 hybrides (testeur x lignées du panel) sur 9 environnements ont été retenues pour les régressions du nombre de grains sur JS et IFT. Ce modèle obtient une moyenne de  $R^2$  ajusté très faible (de l'ordre de 0.26), ce qui indique qu'il n'est pas globalement adapté aux



**Figure II.20 : Relation entre la date de floraison (en jours quantième) et le rendement potentiel ou le coefficient de régression du rendement sur l'indice de stress hydrique**



**Figure II.21 : Relation entre le coefficient de corrélation entre le rendement et la floraison pour 12 variétés communes de blé en fonction des précipitations sur 16 essais.  $R^2=0.62$ . Une réduction des précipitations implique un stress hydrique terminal plus fort et donc un désavantage des variétés ayant une plus longue durée de cycle. (Blum, 2011)**



lignées du panel d'association. Les modèles ne comprenant que JS ou IFT sont encore moins adaptés.

### **Effet des testeurs :**

Une analyse de covariance consistant à expliquer le rendement (grains et huile) en fonction de l'interaction entre l'indicateur de stress et un effet « paire de testeurs » a été menée (effet  $Groupe_k$  dans l'équation qui permet de séparer les environnements en 2 groupes : ceux ayant la paire de testeurs 83HR4gms/FS71501 et ceux avec SOLR0001M/AT0521). Pour le rendement, l'effet « groupe » est significatif pour 115 lignées alors que pour le rendement en huile, il n'est significatif que pour 9 lignées. Par contre, ni pour le rendement grains, ni pour le rendement en huile, la nature du groupe des testeurs n'a d'effet sur la pente, ce qui permet de confirmer que cet index multilocal prend en compte la réponse globale des géotypes indépendamment de la paire de testeur utilisée. Cela ne nous permet pas cependant de conclure qu'il n'y a pas d'effet liée à la différence entre testeurs sur le même environnement (ou/et entre lignées B et R). Cette possibilité pourra être écartée dans la recherche d'association grâce à la correction de cet effet dans les modèles statistiques prenant en compte la structure du panel.

### **Effet de la précocité de floraison :**

Nous avons également vérifié l'impact de la date de floraison sur les différents paramètres de nos régressions multiples. Alors que pour le rendement en huile, aucune tendance n'est clairement visible, une corrélation assez faible se dessine pour le rendement entre la pente de la régression et la date de floraison d'une part et davantage entre le rendement potentiel et la floraison. Cela signifie que la tardiveté de floraison est associée à un meilleur rendement. Ce résultat, décrit à de nombreuses reprises dans la littérature pour différentes espèces cultivées, s'explique par le fait que plus le cycle de développement est long, plus la plante accumule de la biomasse et potentiellement du rendement. Cependant cette relation n'est pas aussi simple, comme en témoigne nos graphiques de corrélations (Figure II.20) où les nuages de points sont peu étirés. En effet en fonction du stress, surtout s'il est tardif, une durée de cycle trop longue peut constituer un handicap, comme illustré sur le blé (Figure II.21).



## Chapitre III Analyse du panel et choix de modèles de génétique d'association

### III.1 Etude bibliographique : la génétique d'association chez les plantes

La majorité des caractères d'intérêts agronomiques sont des caractères quantitatifs dépendants de l'action de nombreux gènes et de leurs interactions avec l'environnement. La compréhension de ces déterminismes génétiques complexes constitue une préoccupation majeure pour les généticiens dans la perspective de l'amélioration des plantes. Depuis l'article fondateur de Lander et Bostein en 1989, qui a été suivi d'une multitude d'analyses visant à cartographier les locus affectant les caractères quantitatifs ("QTL" : Quantitative Trait Locus), de nombreuses avancées ont été réalisées, tant sur le plan des méthodes statistiques que sur celui des technologies de l'ADN. Des millions de marqueurs moléculaires ont ainsi été générés entraînant la densification des cartes génétiques. Face aux approches classiques de détections de QTL dans des populations biparentales (analyses de liaison) l'utilisation de la génétique d'association s'est rapidement imposée comme une méthode de choix. D'abord appliquée à la génétique humaine (Corder *et al.*, 1994), cette méthode a été introduite chez les plantes à partir des années 2000 (Thornberry *et al.*, 2001).

La génétique d'association partage le même principe que l'analyse de liaison : il s'agit de trouver une corrélation statistique entre génotype et phénotype, mais contrairement à l'analyse de liaison qui utilise des familles au pedigree connu (F2, lignées recombinantes...), le matériel étudié rassemble des individus dont l'apparentement n'est pas contrôlé. La diversité présente est ainsi plus large que celle issue de deux parents choisis généralement pour leur différence sur un caractère donné. De plus, le plus récent ancêtre commun à deux individus étant plus éloigné dans le temps, les événements de recombinaisons historiques ont entraîné la cassure du génome en fragments de plus petites tailles que ce qui pourrait être obtenu après seulement quelques générations dans une famille expérimentale. Ainsi, si une association entre un marqueur et un phénotype est détectée, la probabilité d'être proche du polymorphisme causal (résolution) dépend de la taille de ces fragments, c'est à dire de l'étendue du déséquilibre de liaison (DL).

Concept important pour la mise en œuvre de la génétique d'association, le déséquilibre de liaison se définit comme l'association préférentielle de certaines combinaisons alléliques. Plusieurs statistiques utilisées pour mesurer le DL sont décrites dans la table III.1. Leur

Statistique	Définition	Formule	Références
$D$	Déviation entre la fréquence haplotypique observée et attendue	$D = p_{AB} - p_A p_B$	Lewontin and Kojima, 1960
$D'$	Mesure D standardisée entre 0 et 1	$D' = D/D_{max}$ avec $D_{max} =$ $\begin{cases} \min\{p_A p_b, p_a p_B\} & \text{si } D < 0 \\ \min\{p_A p_B, p_a p_b\} & \text{si } D > 0 \end{cases}$	Lewontin, 1964
$r^2$	Carré du coefficient de corrélation de Pearson $r$	$r = \frac{D}{p_A p_a p_B p_b}$	Hill and Robertson, 1968

**Table III.1 : Principales statistiques de mesure du DL.** Soit une paire de locus avec les allèles A et a sur le premier locus et B et b sur le second.  $p_A, p_a, p_B$  et  $p_b$  correspondent respectivement aux fréquences alléliques,  $p_{AB}, p_{aB}, p_{Ab}$  et  $p_{ab}$  correspondent aux fréquences haplotypiques.

objectif, dans le cadre de la génétique d'association, est d'estimer la valeur prédictive d'un locus sur un autre (Erzoz *et al.*, 2007). Si le DL est complet dans une région, tout marqueur localisé dans cette région aura la même valeur prédictive pour une association avec le phénotype. La décroissance du DL en fonction de la distance génétique et physique est un critère classiquement utilisé pour obtenir une estimation moyenne du DL sur l'ensemble du génome et ainsi permettre d'orienter la densité de marquage nécessaire. Plus le DL décroît rapidement, plus la résolution et le nombre nécessaire de marqueurs pour couvrir le génome sont grands.

L'analyse de liaison et la génétique d'association sont basées sur la présence de DL entre un marqueur et le locus causal mais contrairement à la 1<sup>ère</sup> approche où le DL est essentiellement dû à la distance physique entre locus, la génétique d'association, appelée également « LD mapping », utilise des populations où le DL est influencé par de nombreux facteurs notamment évolutifs. La nature du DL est ainsi dépendante du panel d'étude et de la région du génome étudié.

Parmi ces facteurs, le système de reproduction tient un rôle important. Chez les espèces autogames, le DL est maintenu sur de plus longues distances que chez les espèces allogames car les opportunités de recombinaison sont plus faibles étant donné que les individus ont tendance à être davantage homozygotes (Flint-Garcia *et al.*, 2003). C'est le cas chez le riz, *Arabidopsis* ou le blé (Nordborg 2000 ; Zhang *et al.*, 2010) alors que chez des espèces allogames telles que le maïs ou le tournesol, le DL décroît en général plus vite avec la distance (Tenailon *et al.*, 2001 ; Fusari *et al.*, 2008).

Le niveau de domestication présent au sein de la population influence également l'étendue du DL. En effet la domestication, et plus généralement la sélection, est accompagnée d'un goulot d'étranglement qui réduit la taille de la population entraînant une perte des combinaisons alléliques rares (dérive génétique) et donc une réduction drastique de la diversité. La sélection n'est pas un processus aléatoire puisqu'elle vise à sélectionner un phénotype spécifique. Lorsque celui-ci est extrême, la sélection est dite « directionnelle » et modifie l'étendue du DL en changeant les fréquences alléliques aux QTL responsables du caractère sélectionné (Mackay, 2007). En particulier, le phénomène d'hitchhiking augmente le DL autour des locus sélectionnés en y réduisant le polymorphisme (Maynard *et al.*, 1974). Les goulots d'étranglements issus de la domestication ont par exemple provoqué 30 à 40% de réduction de la diversité chez le maïs et de 40 à 50% chez le tournesol (Liu *et al.*, 2005) ce qui se traduit par une étendue du DL sur de plus longue distance. Ainsi chez le maïs, le DL s'étend sur un Kb pour certaines variétés locales (« landraces »), deux kb pour des lignées fixées et jusqu'à plusieurs centaines de Kb pour des variétés commerciales (Jung *et al.*, 2004). Liu *et al.* (2005)

	H0 n'est pas rejetée	H0 est rejetée	
H0 est vraie	pas d'erreur, "vrai positif" (1- $\alpha$ )	erreur de type I, "faux positif" ( $\alpha$ )	→ Seuil de risque
H0 est fausse	erreur de type II, "faux négatif" ( $\beta$ )	Pas d'erreur, "vrai négatif" (1- $\beta$ )	→ Puissance

**Figure III.1 : Principe d'un test d'hypothèse.** Dans le cadre d'un test d'association, l'hypothèse nulle H0 correspond à l'absence d'association entre marqueur et génotype.

ont également montré une décroissance très rapide du DL (< 200pb) pour le tournesol sauvage en comparaison du tournesol cultivé (~1100 pb).

La sélection balancée, également appelée « avantage aux hétérozygotes », favorise les individus porteurs des deux allèles à un SNP. Ce type de sélection en entretenant la persistance d'un polymorphisme, peut également augmenter l'étendue du DL. Chez *Arabidopsis*, Tian *et al.*, (2002) ont ainsi reporté, sur un intervalle de 20Kb comprenant le gène de résistance aux maladies RPS5, la présence simultanée d'un taux élevé de polymorphisme et d'un DL presque complet.

Enfin, la présence d'une structuration dans la population étudiée figure parmi l'une des sources majeures de DL. Lorsque les croisements ont lieu préférentiellement au sein d'une sous-population, les fréquences alléliques ont tendance à diverger entre sous-populations. Le fait de réunir ces sous populations en une seule crée du DL entre locus non liés physiquement. Tout marqueur dont la fréquence allélique varie selon les sous-populations sera associé au phénotype si celui est corrélé à la structure. Ce phénomène doit être pris en compte en génétique d'association puisqu'il induit la présence de faux positifs, c'est-à-dire de fausses associations marqueur-phénotype (Figure III.1).

Pour corriger la présence de faux-positifs, une des premières méthodes proposée a été le contrôle génomique (Devlin 1999) dont le principe est le suivant: un set de marqueurs répartis au hasard sur le génome est choisi pour estimer le degré d'inflation ( $\lambda_{GC}$ ) de l'hypothèse nulle de non association avec le phénotype. Les p-values des tests sont alors corrigées par cette constante. Cependant, la principale limitation de cette méthode vient du fait qu'elle applique la même correction pour tous les marqueurs conduisant à une perte de puissance pour les marqueurs non corrélés au pedigree (Aistle et Balding, 2009). D'autres approches ont ensuite été largement appliquées, telles que l'estimation de la structure et son intégration sous forme de covariable dans des modèles de régression du phénotype sur le génotype décrit par les marqueurs. L'estimation de la structure est classiquement réalisée à partir de l'approche bayésienne implémentée dans le logiciel STRUCTURE (Pritchard *et al.*, 2000). Pour pallier aux temps de calculs très longs requis par ce logiciel, Price *et al.*, (2006) ont ensuite suggérer l'utilisation de l'analyse en composante principale qui permet de réduire les données à quelques dimensions, maximisant chacune la variabilité. Le modèle mixte de Yu *et al.*, (2006) combine l'estimation de la structure (d'après l'approche bayésienne ou l'ACP) à une estimation plus fine de l'apparentement entre paires de lignées (kinship), l'effet polygénique, inclus sous forme de facteur aléatoire.

Le niveau de structuration présent dans une population a une influence directe sur la puissance de détection des caractères (probabilité de détecter un vrai positif, cf. définition





dans la Figure III.1), surtout quand ceux-ci sont corrélés à la structure. Un exemple communément cité concerne la floraison chez *Arabidopsis* (Aranzana *et al.*, 2005), un caractère très corrélé à la structure de par l'histoire adaptative de l'espèce. Les corrections nécessaires dans les tests d'association engendrent la présence de faux négatifs, c'est-à-dire des associations marqueur-caractère non détectées car corrélés à la structure.

La structure n'est pas le seul levier sur la puissance de la génétique d'association. La distribution des fréquences alléliques en est un essentiel car les allèles rares sont très difficiles à détecter (Myles *et al.*, 2009). Le nombre d'allèles rares susceptibles d'avoir un impact sur le phénotype est important et constitue une partie de l'héritabilité « manquante », c'est-à-dire non expliquée par les marqueurs détectés comme associés aux caractères étudiés (Manolio 2009). L'importance de ces allèles rares dépend de l'architecture génétique du caractère étudié : nombre de locus et effet de chacun sur le contrôle du caractère. L'utilisation de l'analyse de liaison, où la fréquence allélique peut artificiellement être augmentée, est souvent invoquée pour pallier ce problème, toutefois elle ne permet pas de détecter des effets faibles. L'utilisation combinée de population biparentale et de panel d'association pour la détection de QTL présente ainsi des avantages intéressants (Cadic *et al.*, 2013).

La puissance est également dépendante de la densité de marqueurs intrinsèque compte tenu du niveau de DL présent. Une densité forte est nécessaire lorsque le DL décroît rapidement. C'est ainsi que les stratégies de génétique d'association ont d'abord été ciblées sur des gènes candidats impliqués dans les mécanismes liés aux caractères étudiés. A moins que la voie métabolique ciblée soit très bien caractérisée, le choix des gènes candidats n'est en général pas un exercice facile, tout le génome pouvant être considéré comme candidat lorsque l'on étudie des métabolismes complexes. De plus, certains gènes candidats ont été identifiés chez des espèces modèles en conditions contrôlées, éloignées des conditions naturelles que l'on peut rencontrer au champ par exemple (Nordborg et Weigel, 2008).

Grâce aux avancées des technologies de séquençage et de génotypage, l'utilisation de la génétique d'association à l'échelle du génome entier (GWAS : genome wide association studies) est devenue possible pour de plus en plus d'espèces. Cette thèse sera l'occasion de comparer les deux approches : gènes candidats et GWAS.

Depuis l'étude pionnière réalisée sur l'association entre le polymorphisme du gène *dwarf8* et la floraison et hauteur de plante chez le maïs (Thornsberry *et al.*, 2001), des études d'association ont été publiées sur de nombreuses espèces, incluant le maïs, le riz, le blé, le sorgho, l'orge, le soja. La table III.2 résume quelques-unes de ces études ainsi que les méthodes utilisées. Des résultats notables ont été obtenus. Par exemple, un échantillon de 192 accessions d'*Arabidopsis* a permis d'identifier de nombreux QTL avec des effets importants

Espèce	Matériel végétal	Caractère	Génotypage	Méthode de recherche d'associations	Référence
<b>Arabidopsis</b>	95 accessions diverses	floraison/résistance aux pathogènes	haplotypes (gènes candidats)	Kruskal-Wallis (caractère quantitatif) et test du $\chi^2$ (caractère binaire)	Aranzana <i>et al.</i> , 2005
	197 accessions	floraison	200 000 SNP	Wilcoxon rank-sum test et MLM	Brachi <i>et al.</i> , 2010
	199 accessions	floraison/défense/développement	250 000 SNP	Wilcoxon rank-sum test (variable qualitative),	Atwell <i>et al.</i> , 2010
	473 accessions	floraison	213 000 SNP	MLM	Li <i>et al.</i> , 2010
<b>Riz</b>	413 cultivars	caractères agronomiques divers	44 000 SNP	MLM	Zhao <i>et al.</i> , 2011
	950 cultivars	floraison et rendement	1.3 millions de SNP	GLM et MLM	Huang <i>et al.</i> , 2011
<b>Maïs</b>	282 lignées	provitamineA	8 gènes candidats	MLM	Harjes <i>et al.</i> , 2008
	553 lignées	acide oléique	8 590 SNP	Kolmogorov-Smirnov test	Belo <i>et al.</i> , 2008
	5000 RIL	architecture foliaire	1.6 million SNP	Modèle multi-SNP	Tian <i>et al.</i> , 2011
	350 lignées	caractères de productivité en conditions	44 000 SNP	GLM	Xue <i>et al.</i> , 2013
<b>Blé</b>	95 cultivars	taille des grains et qualité meunière	62 SSR	MLM	Breseghello <i>et al.</i> , 2005
	189 accessions	caractères de productivité en conditions	179 SSR (gènes candidats)	GLM et MLM	Maccaferri <i>et al.</i> , 2010
	208 lignées élites	germination sur pied	1 166 DaRT et SSR	GLM et MLM	Kulwal <i>et al.</i> , 2012
<b>Orge</b>	500 lignées élites	caractères morphologiques	1 536 SNP	MLM	Cochram <i>et al.</i> , 2010
	185 accessions cultivées et 38	caractères de productivité en conditions	816 marqueurs (SNP, SSR, DaRT)	MLM	Varshney <i>et al.</i> , 2012
	192 accessions	tolérance au froid	1 536 SNP	MLM	Visioni <i>et al.</i> , 2013
<b>Colza</b>	172 accessions diverses	teneur en huile	115 SSR	GLM	Zou <i>et al.</i> , 2010
	128 lignées	nécrose du collet	72 SSR	MLM	Jestin <i>et al.</i> , 2010
<b>Soja</b>	257 cultivars	morphologie de la graine	135	Modèle mixte épistatique	Niu <i>et al.</i> , 2013
	115/137 lignées	chlorose ferrique	24 SSR	MLM	Wang <i>et al.</i> , 2008
<b>Tournesol</b>	94 lignées	résistance au Sclerotinia	haplotypes (16 gènes candidats)	MLM	Fusari <i>et al.</i> , 2012
	271 lignées	floraison et ramification	5 359 SNP	MLM	Mandel <i>et al.</i> , 2013

**Table III.2 : Exemples de publications utilisant des méthodes de génétiques d'association chez les principales espèces cultivées.**

accessions : collection de plantes provenant d'un lieu spécifique, cultivars: sous ensemble distinct d'une espèce, parfois sélectionné, ayant un comportement similaire dans leur environnement d'adaptation. MLM : modèle mixte, GLM : modèle linéaire ; SSR : Simple Sequence Repeats, DaRT : Diversity Arrays Technology, SNP: Single Nucleotide Polymorphisms

(supérieurs à 20% de la variance phénotypique pour 44 des 50 caractères) ainsi que des gènes candidats bien connus, prouvant ainsi le concept (Atwell *et al.*, 2010) malgré une taille de population relativement faible. Une autre étude d'association portant sur un panel de 950 accessions de riz a obtenu un succès similaire (Huang *et al.*, 2010) avec des effets alléliques de l'ordre de 46% pour certains caractères. Le système de reproduction, autogame pour le riz et Arabidopsis, associé à un déterminisme génétique caractérisé par des allèles communs à effets forts a contribué sans doute au succès de ces études. Chez le maïs, le déterminisme génétique semble plus complexe et la présence de nombreux allèles rares constitue une limitation pour la puissance de détection des associations. Cependant, des résultats intéressants ont été obtenus à partir d'approche gènes candidats, par exemple pour un gène (*IcyE*) impliqué dans la synthèse de provitamineA (Harjes *et al.*, 2008) ou encore à partir d'approche « whole genome » avec par exemple l'identification d'un locus ayant un effet majeur sur l'acide oléique et très proche du gène candidat *fad2* (Belo *et al.*, 2008).

Chez le tournesol, l'utilisation de la génétique d'association est très récente. La première étude date de juin 2012 et a permis d'identifier un polymorphisme impliqué dans la résistance au sclérotinia, pathogène très dommageable pour le tournesol à partir d'un panel de 94 lignées et de 16 gènes candidats. Mandel *et al.*, (2013) ont mené une étude plus ambitieuse : à partir d'un panel de 271 accessions diverses de tournesol cultivé (lignées et variétés populations fixées), les auteurs ont identifié plusieurs marqueurs associés à la ramification et la date de floraison. Grâce à l'accessibilité de densités de marquage plus importantes, le tournesol, pour lequel la diversité ainsi que l'étendue du DL ont été bien caractérisées dans différents panels, est devenu un excellent candidat pour l'utilisation de la génétique d'association. Cette méthode offre le potentiel d'identifier du polymorphisme d'intérêt pour l'amélioration des critères agronomiques.

### **III.2 Panel : origine et données moléculaires**

#### *III.2.1* Origine du matériel

Une core collection de 384 lignées a été constituée à partir d'un set de 752 lignées collectées à l'INRA ainsi que dans plusieurs sociétés semencières (Coque *et al.*, 2008). Ce set, comprenant des variétés élites ainsi que des lignées développées par introgression de différents écotypes sauvages du genre *Helianthus* dans du matériel élite, a été génotypé sur 51 microsatellites (SSR) correctement répartis sur le génome (3 par groupe de liaison). Ces

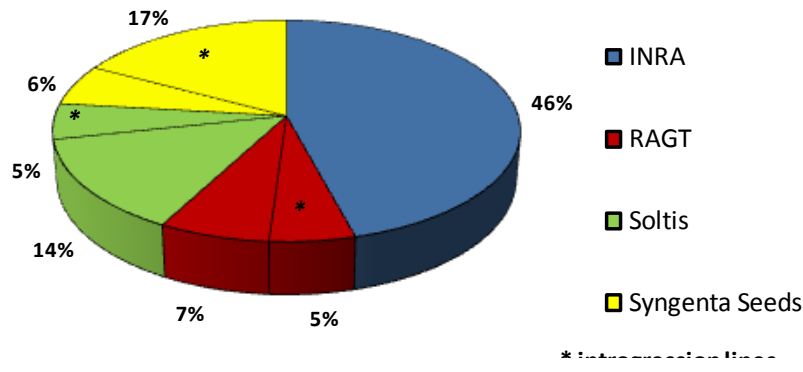


Figure III.2 : Entreprises d'origine des lignées de la core collections. Les étoiles représentent les lignées résultant d'introgessions par du matériel sauvage

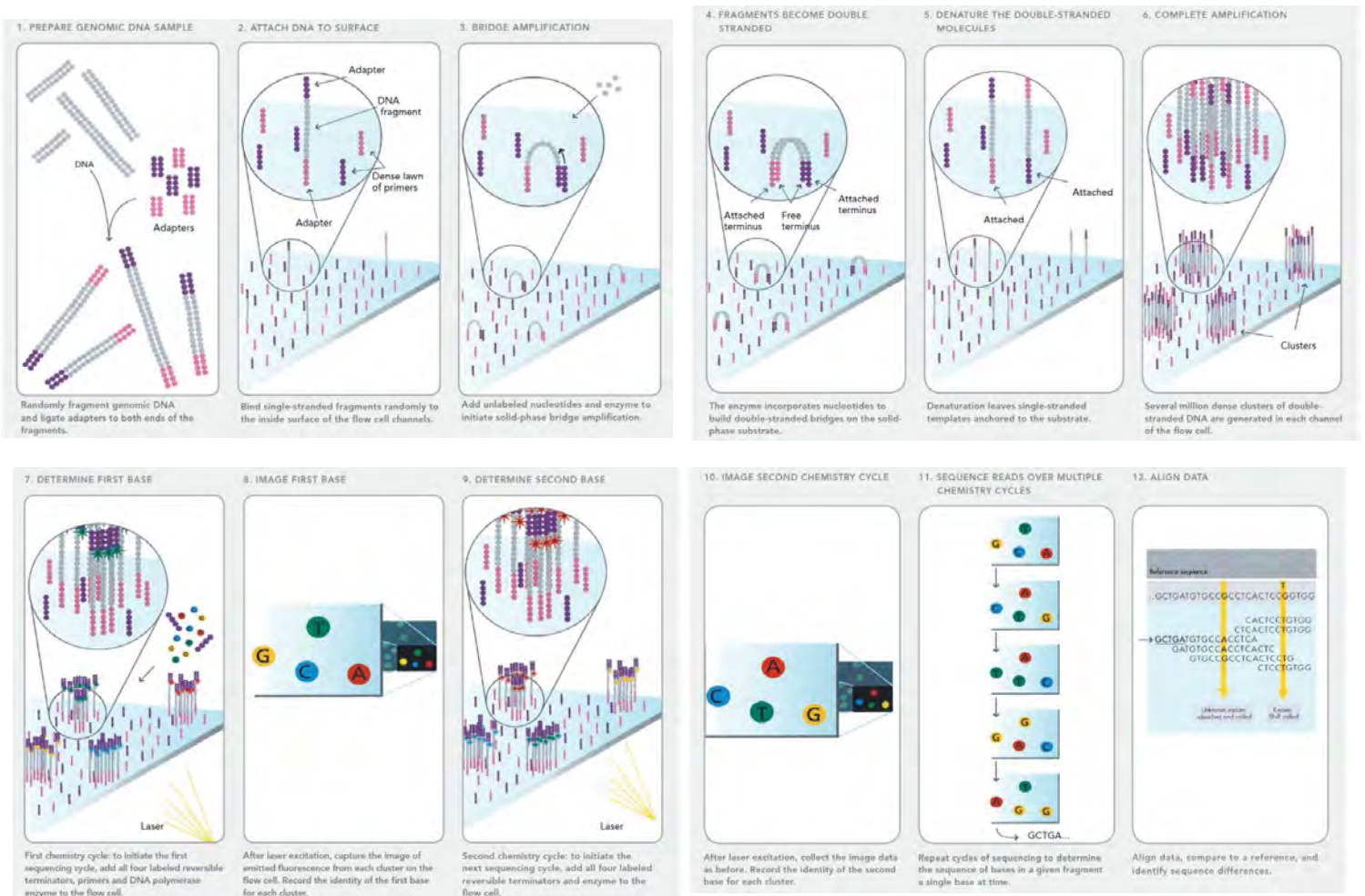


Figure III.3 : Principe de la technologie de séquençage par synthèse d'Illumina (Genome Analyser)

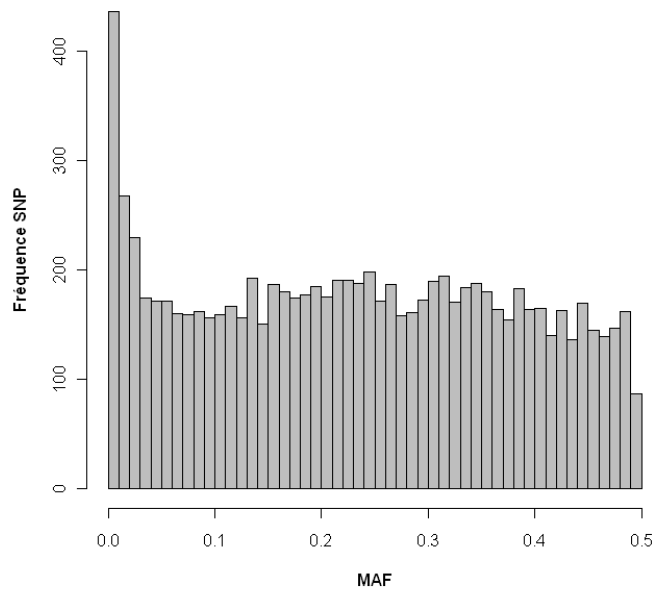
données moléculaires ont permis à Coque *et al.*, de sélectionner un ensemble de lignées représentant 100% de la diversité du panel initial. Pour cela, les auteurs ont choisi 12 lignées diverses et présentant un intérêt pour différentes applications (certaines étant des parents de populations de références) afin de former un noyau autour duquel plusieurs core collections ont été emboîtées en maximisant la diversité allélique à chaque étape. Même si la diversité maximale du set initial a été atteinte avec 220 lignées, la collection a été étendue à 384 lignées afin d'augmenter la puissance des tests d'association. Le panel final est ainsi composé d'accessions publiques fournies par l'INRA et d'accessions privées fournies par trois entreprises semencières, chacune ayant apporté du matériel cultivé et des lignées d'introgessions (Figure III.2).

### III.2.2 Données moléculaires

Le polymorphisme de type SNP a été identifié en utilisant trois approches différentes. La première approche basée sur un séquençage génomique de type Sanger a permis de séquencer dix lignées sur un ensemble d'amplicons non ciblés en exploitant les séquences EST disponibles dans les banques de séquence publique.

La seconde approche a consisté à séquencer le transcriptome de huit lignées en utilisant la technologie Solexa (Illumina). Une librairie de fragments de cDNA représentatifs du transcriptome a été obtenue par cassure mécanique et clusterisé sur un support solide. Cette technologie de « séquençage par synthèse » permet de conduire des réactions de séquençage parallèlement sur ces clusters, la totalité des séquences générées (30 à 40 pb en moyenne) étant ensuite assemblées en une séquence consensus (Figure III.3). Enfin, la troisième approche s'est appuyée sur le séquençage génomique (Solexa) de gènes candidats impliqués chez *Arabidopsis* dans la signalisation hormonale, le développement, la réponse à des stress abiotiques ou biotiques ainsi que de facteurs de transcription, sur un sous ensemble de 48 lignées. Les lignées séquencées à travers chaque approche ont été choisies de manière à être bien réparties dans la core collection. Alors que les deux premières approches visent à obtenir un grand nombre de SNP sur l'ensemble du génome, avec un coût toutefois plus important pour la méthode Sanger, maintenant supplantée par les NGS (Next generations sequencing technologies), l'approche « gène candidat » est davantage qualitative.

Parmi l'ensemble des SNP découverts, 12136 ont été génotypés sur l'ensemble de la core collection en utilisant différentes plateformes (BeadXpress et Infinium). 8844 SNP polymorphes ont été validés après ajustement pour chaque SNP des positions des clusters



**Figure III.4 : distribution des MAF sur les 8844 SNP géotypés**

(génotypes AA, AB ou BB) dans le logiciel Genome Studio d'Illumina, utilisé pour convertir les intensités de fluorescence en génotype ([http://www.illumina.com/software/genomestudio\\_software.ilmn](http://www.illumina.com/software/genomestudio_software.ilmn)). Parmi ces 8844 SNP, 61 % proviennent de l'approche de séquençage du transcriptome, 15 % du séquençage génomique non ciblé et 24 % du séquençage génomique ciblé sur des gènes candidats. Alors que 64 % des gènes candidats contiennent plus de un SNP génotypé par gène, 91 % des gènes non ciblés (approches 1 et 2) ne contiennent qu'un SNP génotypé par gène.

80 lignées de la core collection ont dû être éliminées suite à des résultats de génotypage non cohérents avec des données antérieures amenant le panel d'association au nombre de 304 lignées, toutes ayant un taux d'hétérozygotie inférieur à 10%. La fréquence allélique minimum par SNP (MAF) s'élève en moyenne à 0.23 (Figure III.4).

Parmi les 8844 SNP, 6507 ont été positionnés sur une carte consensus appartenant à Biogemma, construite à partir de quatre populations de lignées recombinantes dont la population INEDI (cf. Chapitre V), utilisée comme pivot, et de deux populations de F2. Cette carte a une densité d'environ un marqueur tous les 0.50 cM.

### **III.3 Structuration du panel**

#### *III.3.1* Matériels et méthodes

Afin d'analyser la structuration du panel, les SNP ont été sélectionnés sur leur pourcentage de données manquantes (inférieurs à 10%) et sur leur MAF. De plus, les SNP appartenant à des gènes candidats ont été écartés. Ce choix est important notamment pour l'utilisation du logiciel STRUCTURE (Pritchard, 2000) qui suppose que chaque marqueur est en équilibre d'Hardy-Weinberg et donc non soumis à des pressions de sélection.

La figure III.4 qui représente la distribution des SNP selon leur MAF montre une nette inflation du nombre de SNP ayant une MAF inférieure à 1 %. Nous avons donc choisi ce seuil pour trier les SNP utilisés pour la structure, ce qui nous amène à conserver 5679 SNP au total. La version 2.2 du logiciel STRUCTURE a été utilisée pour inférer la structure du panel en spécifiant un modèle avec admixture et fréquences alléliques corrélées (Falush *et al.*, 2003). Dix répétitions pour un nombre de groupe (Nk) variant de un à 11 ont été lancées avec un « burn in time » de 50 000 et 100 000 itérations. Les critères d'Evanno (Evanno *et al.*, 2005) ont aidé à l'identification du nombre le plus probable de sous-populations. Le principe de ces critères consiste à tracer le maximum de vraisemblance en fonction de Nk et de choisir la valeur de Nk pour laquelle le changement de pente est le plus important. Les valeurs de  $F_{st}$

marqueur	locus	alleles	alleles/locus	PIC
SSR	48	410	8.5 (3-19)	0.64 (0.19-0.89)
SNP	5923	11846	2	0.35 (0.02-0.05)

**Table III.3 : Comparaison de quelques paramètres entre les deux types de marqueurs moléculaires utilisés pour la structuration.** Le PIC (Polymorphism Information Content) traduit l'utilité des marqueurs pour la découverte de QTL.

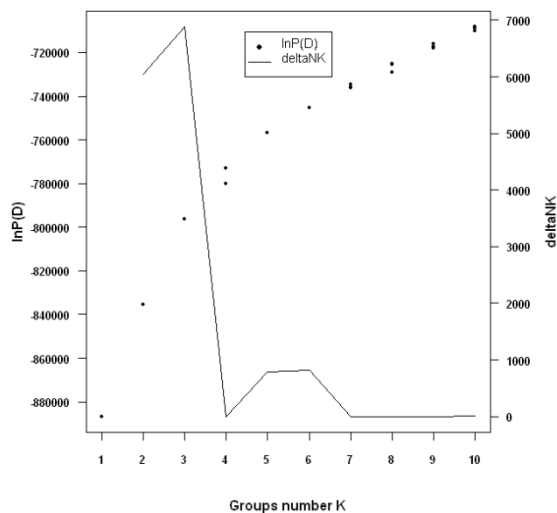


fournies par le logiciel, analogues à la statistique de différenciation entre populations  $F_{st}$  de Wright, ont permis d'estimer la divergence inter groupes.

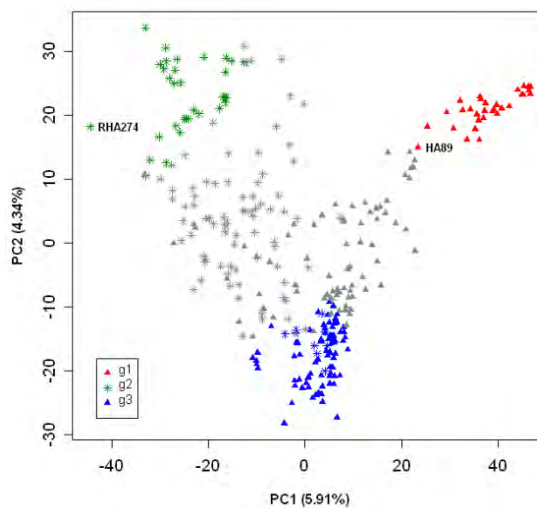
En complément de l'approche bayésienne, une analyse en composante principale a été menée sur le même set de SNP après normalisation de la matrice de génotypage selon Patterson *et al.*, (2006). C'est également suivant les recommandations de cet auteur que nous avons évalué le nombre de composantes principales statistiquement significatives. Les valeurs propres issues de la décomposition de la matrice variance covariance suivent une loi de Tracy-Widom, les probabilités pour chaque composante d'être significative ont été calculées grâce au package RMTstat (Johnstone *et al.*, 2009).

L'apparentement entre toutes les paires de génotypes a été estimé à partir des SNP filtrés sur 10% de données manquantes mais pas sur les MAF (au final 5904 SNP) en utilisant l' AIS (Alikeness in State) disponible dans le logiciel Cocoa (Maenhout *et al.*, 2009). Cet estimateur correspond à la probabilité qu'ont deux individus de posséder le même allèle à un locus. Si ces allèles semblent être les mêmes, ils ne dérivent pas forcément d'un ancêtre commun, ce qui introduit un biais dans l'estimation. Cependant, la matrice d'apparentement qui en découle (que nous appellerons Kais) est définie semi-positive (en absence de donnée manquante) et donc directement utilisable dans les modèles de génétique d'association.

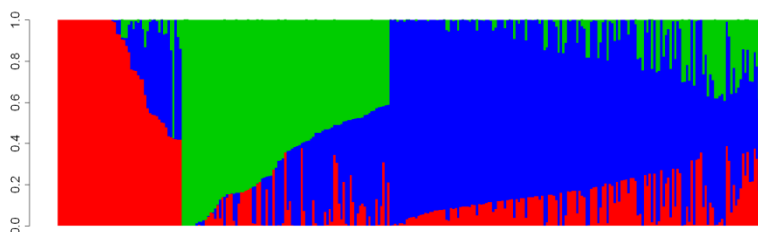
En comparaison de ces analyses basées sur un set de SNP, nous avons examiné la structure en utilisant les données microsatellites ayant servi à la construction du panel. Parmi les 51 SSR bien répartis sur le génome, nous en avons éliminé deux qui présentaient un taux de données manquantes supérieur à 49%. Cependant, il reste 20 marqueurs ayant plus de 30% de données manquantes et la MAF moyenne sur l'ensemble des marqueurs s'élève à 0.52. Les caractéristiques de ces marqueurs en comparaison des SNP sont présentées dans la table III.3. Nous n'avons pas appliqué les mêmes critères de sélection (données manquantes et MAF car les marqueurs SSR sont beaucoup plus fiables et les ambiguïtés de typage sont le plus souvent à l'origine des données manquantes. Les SSR révèlent un caractère plus informatif, d'après les valeurs de PIC (Table III.3), que les SNP mais leur nombre reste cependant très faible comparé aux SNP disponibles. Les mêmes méthodes ont été appliquées sauf pour l'ACP où les données n'ont pas été normalisées car, d'après Patterson *et al.*, (2006), cette étape n'est pas justifiée dans le cas des données microsatellites.



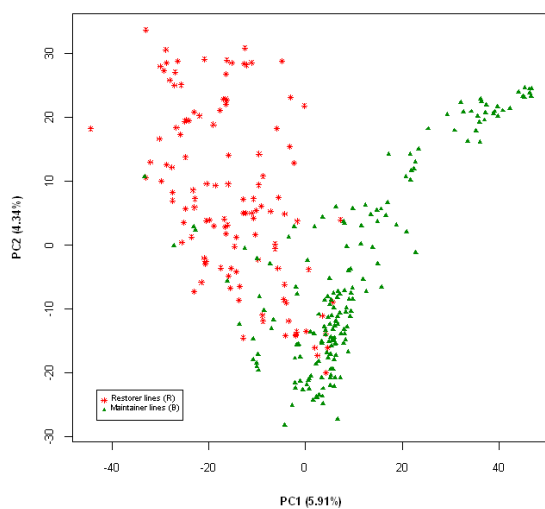
**a**



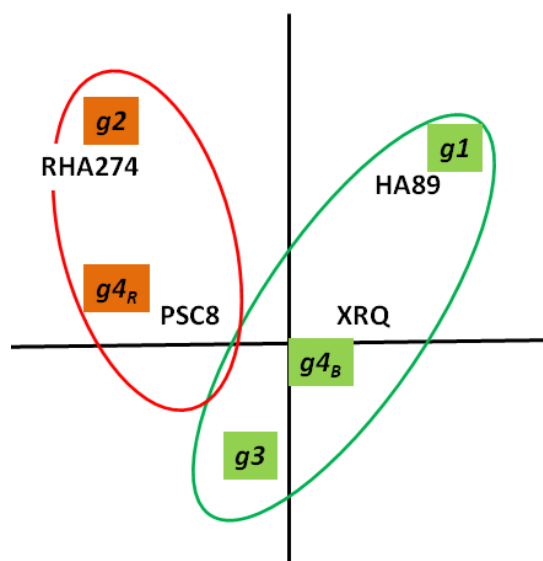
**b**



**c**



**d**



**e**

**Figure III.5 : Résultats de la structuration réalisée à partir des SNP :** distribution de la vraisemblance et du critère delta d'Evanno pour un nombre de groupe de 1 à 10 (a), premier plan de l'ACP avec une coloration selon le groupe inféré par le logiciel STRUCTURE (b), résultats de STRUCTURE (c), ACP avec coloration selon le type B/R des lignées (d) et résumé de la structure du panel (e).

### III.3.2 Résultats et discussion

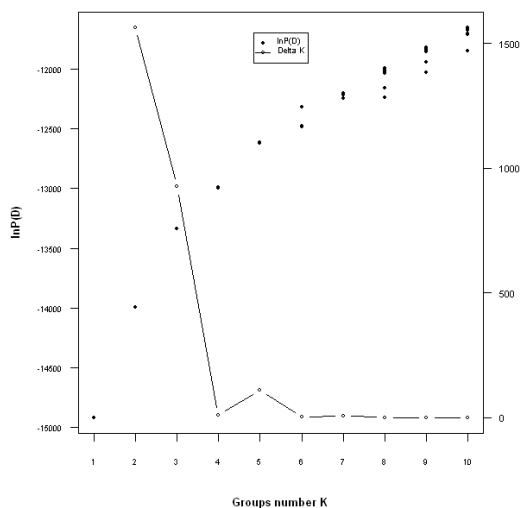
#### III.3.2.1 Inférence de la structure à partir des SNP

Concernant les résultats issus du logiciel STRUCTURE, la distribution de la vraisemblance moyenne en fonction de  $N_k$  n'atteint pas de plateau mais laisse apercevoir une certaine stabilité entre les répétitions (figure III.5.a). Le deuxième critère d'Evanno  $\Delta k$ , basé sur la dérivée au second ordre du logarithme de la vraisemblance bayésienne  $\ln P(D)$ , indique clairement un nombre de groupes le plus probable à trois (tandis que  $N_k=5$  et  $N_k=6$  ressortent plus faiblement). Les individus affectés à l'un de ces trois groupes ( $g_1$ ,  $g_2$  et  $g_3$ ), avec un pourcentage de leur génome assigné supérieur à 80, apparaissent en couleur dans l'ACP (Figure III.5.b). Le reste du panel, que nous désignerons «  $g_4$  » (en gris sur la figure) rassemble les individus qui n'ont pas pu être affectés à un groupe. Les résultats de l'ACP sont très similaires à ceux de STRUCTURE. D'après le test de Tracy Widom, les trois premières composantes principales, qui représentent un total de 13.21 % de la variabilité, sont statistiquement significatives.

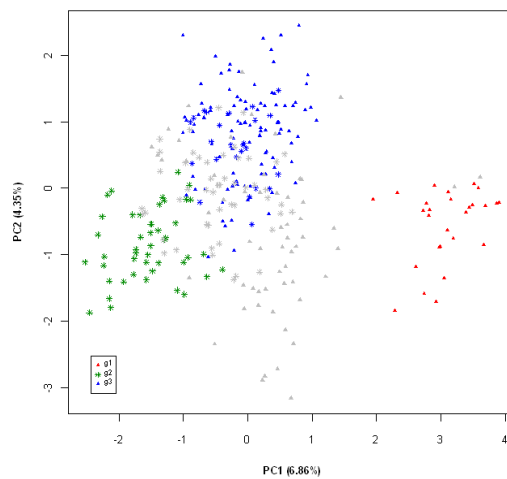
Les groupes  $g_1$  et  $g_2$ , qui apparaissent respectivement en rouge et vert sur les figures, correspondent à des lignées d'introgressions fournies par l'une des entreprises semencière, Syngenta Seeds,  $g_1$  correspondant aux lignées B (mainteneuse de la stérilité) et  $g_2$  aux lignées R (restauratrices de la fertilité). La plupart des lignées d'introgressions B de Syngenta seeds sont assignées à  $g_1$  à plus de 98%, alors que 22 lignées d'introgressions R de Syngenta sur 36 sont rattachées à  $g_2$  à hauteur de 90%. Le groupe  $g_3$  contient principalement des lignées B publiques fournies par l'INRA.

D'après les valeurs de  $F_{st}$  et les taux d'admixture de STRUCTURE, ces trois groupes ont des niveaux de divergence différents. Les groupes  $g_1$  et  $g_2$  ( $F_{st}=0.57$  et  $0.40$ ) apparaissent les plus divergents du reste du panel, probablement en raison de la contribution des écotypes sauvages. Au contraire les groupes de lignées d'introgressions fournies par les autres entreprises ne présentent pas de répartition différenciée au sein du panel, peut être car les parents récurrents choisis pour les backcross étaient d'origines diverses.

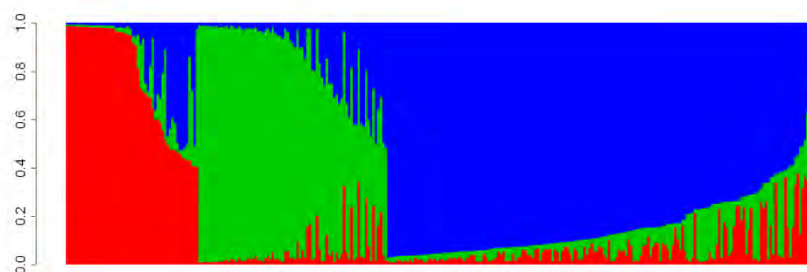
Le groupe  $g_3$  des lignées B INRA est quant à lui peu divergent ( $F_{st}=0.03$ ). De plus les lignées publiques B qui le constituent présentent beaucoup moins de diversité que les lignées publiques R réparties au sein de  $g_4$ . Une hypothèse a pu être formulée pour expliquer ce constat (Figure III.5.e). L'une des lignées R (RHA274) considérée comme fondatrice dans les programmes de sélection car apportant la première à la fois une source de restauration de la



**a**

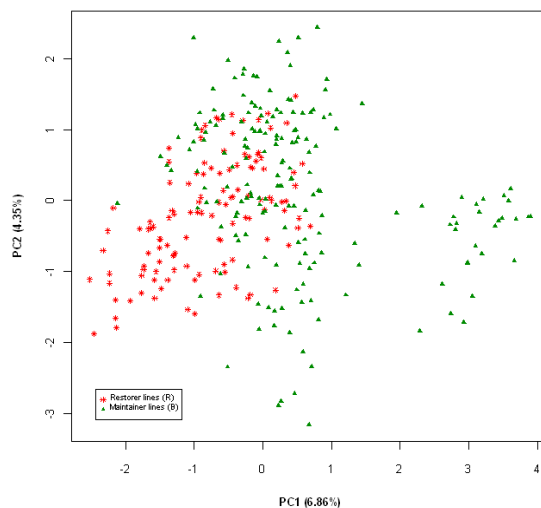


**b**



**c**

**d**



**Figure III.6 : Résultats de la structuration réalisée à partir des SSR :** distribution de la vraisemblance et du critère delta d'Evanno pour un nombre de groupe de 1 à 10 (a), premier plan de l'ACP avec coloration selon le groupe inféré par le logiciel STRUCTURE (b), résultats de STRUCTURE (c), premier plan de l'ACP avec coloration selon le type B/R des lignées

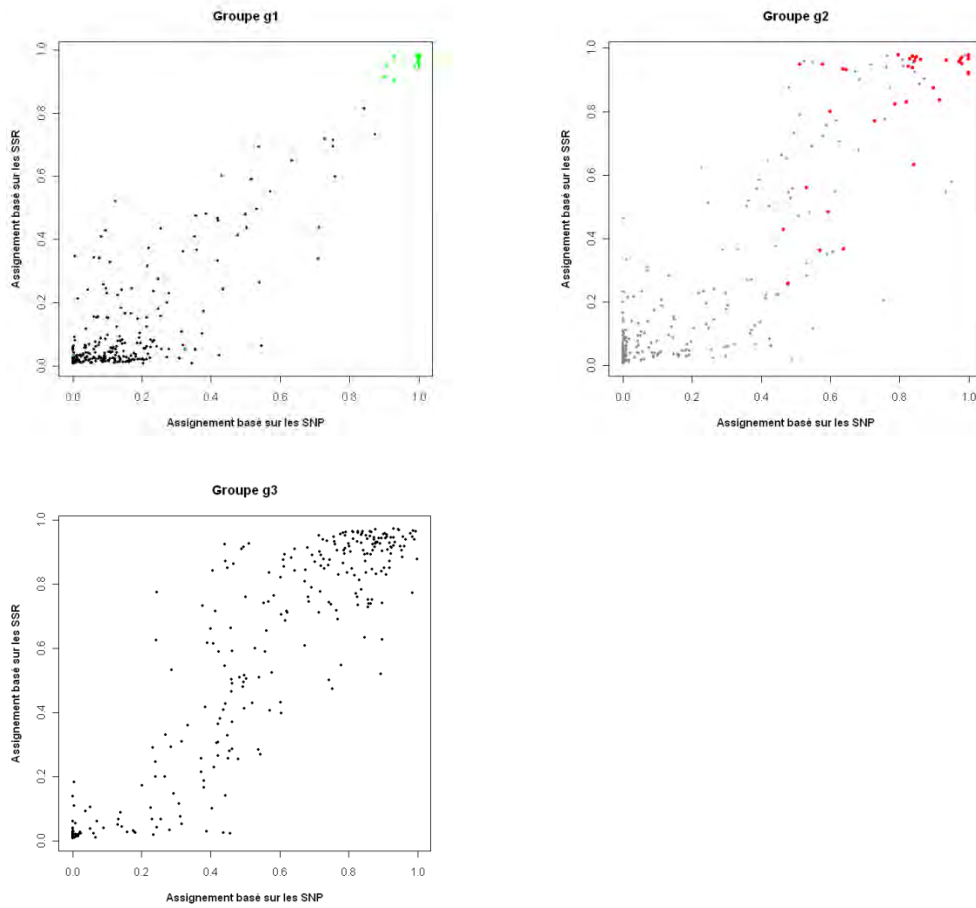
fertilité (locus Rf1) et un gène dominant de résistance au mildiou, apparaît très proche de g2 dans l'ACP même si ce n'est pas une lignée d'introgession. Les lignées R issues du croisement entre RHA274 et les lignées B se dispersent ainsi entre la lignée en question et le groupe g3. De plus l'apport de nouveau matériel étant davantage ciblé sur les lignées R que B, ces dernières étant principalement croisées entre elles, a pu introduire d'avantage de variabilité au sein des R. La plupart des lignées B sont originaires de variétés-population russes. L'une d'entre elle, HA89, également parent fondateur comme RHA274, apparaît relativement distante du groupe g3 des lignées B INRA et plus proche de g1, groupe des lignées d'introgession B de Syngenta. Cette observation suggère que le groupe g3 possède un autre type de fondateur, qui d'après les informations de pedigree pourrait provenir d'Europe de l'Est ou d'introgessions avec du matériel sauvage différent de celles contenues dans g1 et g2.

Les résultats précédents nous ont permis de comprendre davantage les événements qui ont pu être à l'origine de la diversité que nous observons aujourd'hui sur notre panel. Une des caractéristiques principales qui ne ressort pas des résultats de STRUCTURE mais plutôt de l'ACP, concerne la distinction entre lignées R et lignées B. L'ACP souligne cette distinction sur son 1er axe, qui explique 5.91% de la variabilité, en séparant les lignées R à gauche et les lignées B à droite.

La séparation entre pool B et pool R a déjà été mentionné dans la littérature (Berry *et al.*, 1994; Gentzbittel *et al.*, 1994; Hongtrakul *et al.*, 1997) et peut refléter le processus de sélection réciproque qui a conduit à une certaine divergence. Cependant ces deux pools ne forment pas des groupes hétérotiques dans le sens formel, car beaucoup de diversité au sein de chaque groupe persiste.

### III.3.2.2 Comparaison avec les résultats obtenus à partir des SSR

Concernant les résultats obtenus avec les SSR, la vraisemblance n'atteint pas de plateau (Figure III.6.a). On peut noter que les répétitions deviennent plus instables qu'avec les SNP à partir de  $N_k=6$ . Le critère  $\Delta k$  d'Evanno indique une probable partition du panel en deux sous populations : la première correspond au même groupe g1 détecté à partir des SNP et la seconde regroupe 227 individus du panel assignés avec une proportion de leur génome supérieure à 0.80, soit 103 individus de plus qu'avec les SNP pour  $N_k=2$ .

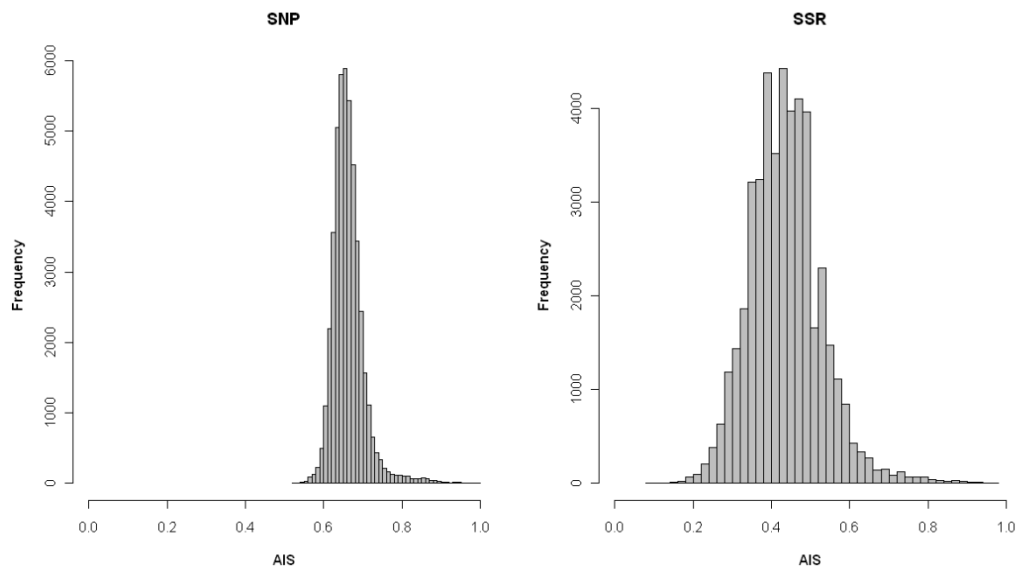


**Figure III.7 : Pourcentage du génome attribué à chacun des groupes à partir des SNP ou des SSR. Les individus en couleur correspondent aux groupes « a priori » des lignées d'introgession B ou R de Syngenta.**

Le critère  $\Delta k$  étant toujours élevé à  $N_k=3$ , nous avons comparé les résultats de STRUCTURE pour ce nombre de groupe. L'ACP (Figure III.6.b) issue des données SSR représente les individus assignés pour chacun des trois groupes. La relation entre les pourcentages de génome attribués aux groupes obtenus à partir des SNP et ceux obtenus à partir des SSR est représenté sur la figure III.7. Le groupe g1 est similaire entre SNP et SSR comme en atteste la figure III.7.a qui montre que toutes les lignées B d'introgessions de Syngenta Seeds (en vert) sont bien affectées à g1, quel que soit le type de marqueurs. Par contre, plus d'individus sont attribués au groupe g2 (Figure III.7.b) en utilisant les SSR, et en particulier plus de lignées R d'introgessions fournies par Syngenta Seeds (en rouge). Ces observations sont cohérentes avec les graphiques d'admixture de STRUCTURE (Figures III.5.c et III.6.c). Quant à g3, les SSR entraînent l'assignation de 32 individus de plus qu'avec les SNP dont des lignées R. Ce groupe apparaît également moins condensé dans l'ACP. De plus, la distinction entre lignées R et B visible sur la 1<sup>ère</sup> PC (Figure III.6.d) est moins claire que sur l'ACP obtenue à partir des SNP. La corrélation entre les 2 types de marqueurs sur cette PC1 s'élève à 0.85 alors qu'elle n'est que de 0.69 sur la 2<sup>ème</sup> PC et 0.55 sur la 3<sup>ème</sup> PC. Cependant le test de Tracy Widom permet de retenir le même nombre de composantes qu'avec les SNP, soit trois.

Le choix des marqueurs peut avoir un impact important sur l'étude de la structuration d'une population et le calcul des distances génétiques séparant ses individus. La comparaison entre SNP et SSR a fait l'objet de plusieurs études, notamment chez le maïs. Hamblin *et al.*, (2007) ont évalué l'utilisation de 89 SSR comparé à 847 SNP pour étudier la structure d'un panel de 259 lignées de maïs. Ils ont observé un pourcentage d'attribution aux sous populations plus faible avec les SNP qu'avec les SSR, ce qu'ils ont interprété comme un manque d'information contenu dans les SNP pour résoudre les relations complexes au sein de leur panel. Van Inghelandt *et al.*, (2010) arrivent aux mêmes conclusions à partir de la comparaison de 359 SSR contre 8244 SNP pour évaluer la structuration d'un panel de 1537 lignées de maïs. Dans tous les cas, les résultats sont globalement cohérents, en termes de groupes détectés entre les deux types de marqueurs.

Dans notre étude, l'utilisation de 48 SSR ou 5923 SNP aboutissent aux mêmes groupes mais avec davantage de lignées assignées dans le cas des SSR que des SNP (pour  $N_k=3$ ). Cette observation va dans le sens des études décrites précédemment mais reste cependant très surprenante étant donné le déséquilibre dans le nombre de marqueurs utilisé : environ 120 fois plus de SNP que de SSR. Van Inghelandt a proposé un nombre de SNP sept à 11 fois supérieur aux SSR pour bien estimer la structuration. Etant bien au-dessus du seuil requis, il



**Figure III.8 : Distributions de l'apparement entre les lignées réalisées à partir des SNP et des SSR.**



semble que d'autres paramètres que le nombre de marqueurs pourraient expliquer ces résultats. L'un des plus variables entre les deux sets de marqueurs est la fréquence des allèles rares, bien plus importante pour les SSR. Ces allèles rares auraient pu être perdus dans la sélection des SNP utilisés pour la structuration.

#### - Apparentement

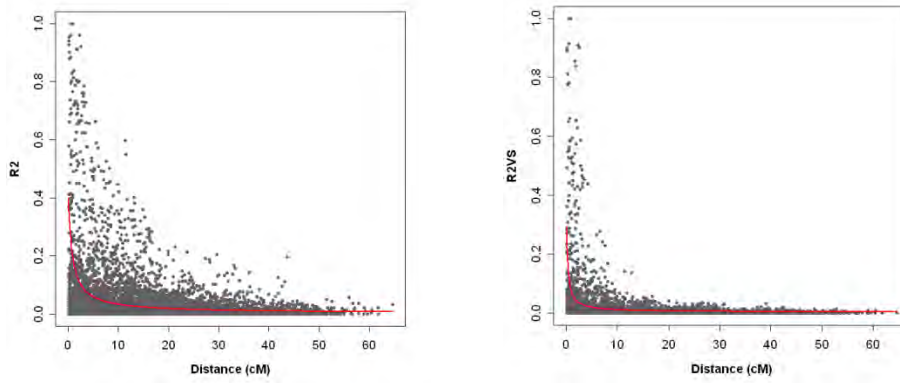
A partir des données SNP, le coefficient AIS moyen entre paires de lignées s'élève à 0.66, suggérant un niveau d'apparentement relativement élevé au sein du panel.

Alors que l'AIS s'étend de 0.52 à 0.99 pour les SNP, il varie plus largement entre 0.09 à 0.97 pour les SSR (Figure III.8), ce qui est en accord avec les résultats de Hamblin et al (2007), Jones *et al.*, (2007) et Van Inghelandt *et al.*, (2010). Parmi les explications proposées par ces auteurs, la différence de fréquence allélique entre les 2 types de marqueurs joue un rôle essentiel. En effet, la présence d'allèles rares chez les SSR aurait tendance à augmenter la proportion de marqueurs polymorphes entre deux individus et ainsi les éloigner.

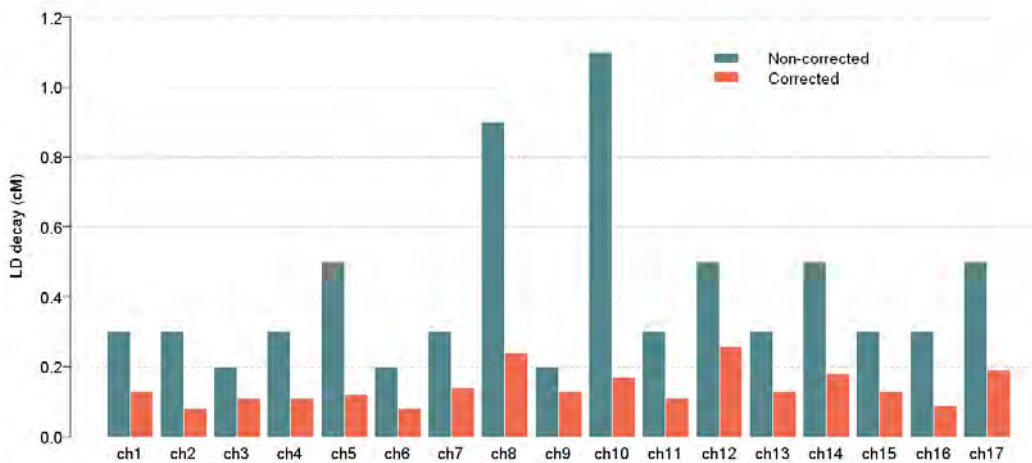
### III.4 Déséquilibre de liaison

#### III.4.1 Matériels et Méthodes

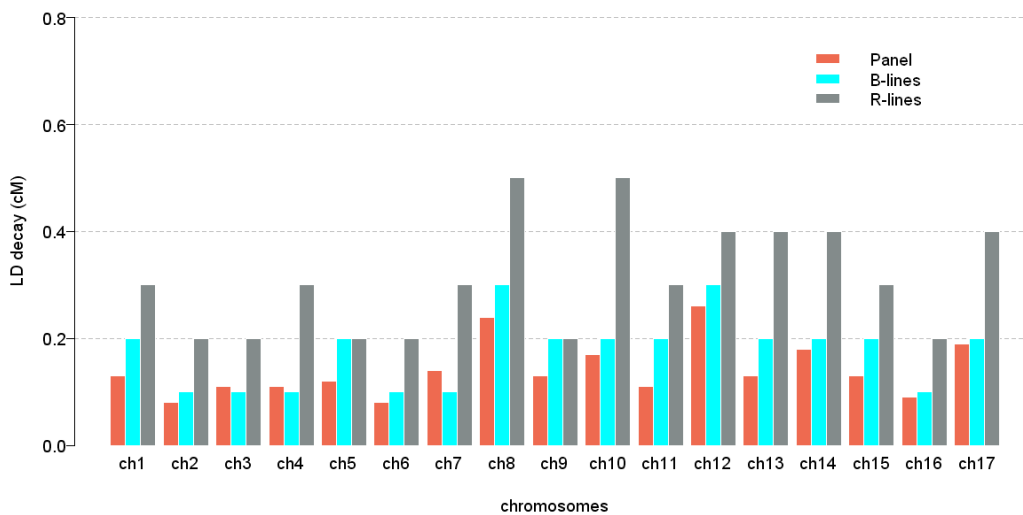
Afin d'estimer l'étendue du déséquilibre de liaison au sein du panel, un set de 1874 SNP génotypés sur le panel et cartographiés sur la carte consensus de Biogemma à des positions toutes différentes, a été sélectionné. Ces SNP contiennent moins de 10% de données manquantes et une MAF supérieure à 5%. Une des statistiques couramment utilisée pour le calcul du DL est le  $r^2$ , équivalent au coefficient de corrélations entre deux marqueurs (cf. section III.1). Elle est en général préférée à d'autres statistiques, notamment pour évaluer le niveau de résolution en génétique d'association, car elle indique directement la corrélation entre le marqueur et le QTL d'intérêt (Flint-Garcia *et al.*, 2003). Après avoir calculé le  $r^2$  pour chaque paire de marqueurs par chromosome, il est d'usage de représenter graphiquement ces valeurs en fonction des distances. Nous avons utilisé la fonction de Hill et Weir (1988) pour modéliser la décroissance du DL en fonction de la distance génétique. Pour déterminer l'étendue moyenne du DL pour chaque chromosome, nous avons choisi un seuil de  $r^2$  égal à 0.2 comme valeur de DL minimale pour détecter une association. Cette valeur, reportée sur la



**Figure III.9 : R2 non corrigé et corrigé en fonction des distances sur la carte génétique**



**Figure III.10 : Etendue du DL moyen corrigé ou non de l'apparentement pour chaque chromosome**



**Figure III.11 : Etendue du DL moyen non corrigé sur le panel, le pool de lignées B et celui de lignées R**

courbe de Hill et Weir, permet d'estimer l'étendue du DL correspondant à la distance en cM sur la carte génétique.

La structuration au sein d'une population peut entraîner la présence de DL entre locus non liés, et provoquer un biais dans l'estimation du DL. C'est pour cela que nous avons utilisé une nouvelle mesure de  $r^2$  corrigé de la structure et de l'apparentement :  $r_{VS}^2$  (Mangin *et al.*, 2011). La matrice d'apparentement correspond au coefficient d'AIS entre toutes les paires d'individus créée à partir du set de 5904 marqueurs (décrit plus haut). Quant à la structure, nous avons choisi un vecteur 0-1 pour le type de lignée (B ou R). La pertinence de ce choix sera justifiée dans la section suivante (choix de modèles). L'étendue de DL a été également évaluée au sein de chaque pool de lignées : lignées B (183 accessions) et lignées R (121 accessions) en utilisant la statistique  $r_{VS}^2$ .

### III.4.2 Résultats et discussion

L'étendue du DL moyen sur l'ensemble des chromosomes varie de 0.41 cM (sans corrections) à 0.14 cM avec correction par la structure et l'apparentement. Ce résultat rappelle l'importance de la structure du panel et sa tendance à augmenter le  $r^2$  entre marqueurs éloignés (Figure III.9). Les valeurs estimées de l'étendue du DL sont cependant très variables d'un chromosome à l'autre (de 0.08 à 0.26). En particulier, deux chromosomes, le LG10 et LG08, se distinguent par un DL décroissant très lentement lorsqu'aucune correction n'est appliquée (Figure III.10). Comme le montre la figure III.11, l'étendue du DL est plus grande pour le pool des lignées R que pour le pool des lignées B. C'est le cas pour la plupart des chromosomes dont le LG08 et LG10. Une des explications plausibles concernant le LG10 pourrait provenir du gène de ramification cartographié sur ce chromosome (Tang *et al.*, 2002). La ramification chez le tournesol augmente le nombre de capitules avec une diminution de leur diamètre, ce qui entraîne une baisse du poids de mille grains mais une augmentation de la teneur en huile (Bachlava, 2010). Ce caractère, *a priori* non désirable en production, a été introduit sous la forme d'un gène récessif chez les lignées R, parents mâles des hybrides, de manière à étendre la période de pollinisation dans les parcelles de production de semences hybrides. Une sélection positive pour ce caractère chez les lignées R a donc pu entraîner l'étendue du DL autour de ce locus, c'est-à-dire un processus d'hitchiking (Meynard *et al.*, 1974). La sélection directionnelle change les fréquences alléliques au QTL responsable du caractère sélectionné mais aussi sur des marqueurs proches (Mackay *et al.*, 2007). Quant au LG08, un cluster de gènes de résistance au mildiou est cartographié sur ce chromosome



(Bouzidi *et al.*, 2002). Les lignées R ayant été fortement sélectionnées sur ce cluster entre 1975 et 1995, peuvent avoir subi le même processus de hitchhiking sur ce chromosome.

Quel que soit le chromosome, l'étendue du DL estimée au sein de chaque pool de lignées (B ou R) demeure plus importante que sur l'ensemble du panel (Figure III.11), avec une étendue plus forte pour le pool R.

L'estimation du DL intra pool ayant été corrigé de l'apparentement entre paires de lignées, il est peu probable qu'une structure sous-jacente soit responsable de cette augmentation du DL.

D'autres hypothèses peuvent être formulées :

- Un biais dû à la taille de l'échantillon a pu être introduit. Yan *et al.*, (2009) ont montré que le DL était influencé par la taille de l'échantillon, avec une décroissance plus lente pour des échantillons de taille faible, ce qui va dans le sens de notre étude (61 lignées en plus dans le pool B par rapport au pool R).
- Invargsson *et al.* (2005) indiquent que les estimations au sein des sous-populations seraient soumises à des erreurs standards plus importantes dues au plus faible nombre de sites polymorphes.
- De plus, l'histoire des recombinaisons dans chaque pool n'a probablement pas été la même étant donné que les pressions de sélection y ont été différentes avec sans doute, du fait de l'historique du développement des hybrides, des pressions plus fortes pour le pool des R, ce qui pourrait justifier que le DL y décroît moins vite.

La valeur moyenne de l'étendue du DL corrigé de la structure et de l'apparentement sur l'ensemble du panel est de 0.14 cM. Pour un génome de l'ordre de 3.5 Gb et une carte génétique de 1800 cM, cette distance génétique correspond environ à une distance physique de 272 Kb. Ce résultat n'est pas surprenant sachant que le panel est constitué d'une part importante de lignées élites. Chez le maïs, des valeurs de 100 à 500 kb ont été observées sur des lignées commerciales (Jung *et al.*, 2004). En tournesol, plusieurs auteurs ont examiné l'étendue du DL sur divers échantillons d'individus et gènes. Kolkman *et al.*, (2007) ont analysé un set de 10 lignées élites et deux accessions sauvages. Ils ont estimé que le DL s'étendait sur 5.5 kb pour un seuil de  $r$  de 0.32. Fusari *et al.*, (2008) font état d'une distance de 643 pb pour un seuil de  $r^2$  de 0.64. Ces études considèrent le DL sur de courtes distances (inférieures à un kb). Il est donc difficile de comparer avec notre étude basée sur l'ensemble du génome. De plus, l'estimation du DL est dépendante de l'échantillon et de la région du génome ciblée, ce qui rend les estimations moyennes peu instructives, à part pour orienter le génotypage dans le cadre de la mise en place de la génétique d'association. Il ressort de nos résultats qu'environ 12 857 SNP seraient nécessaires pour couvrir correctement le génome. Même si dans le cadre de cette thèse, nous n'avons environ que la moitié des SNP requis



disponible, les récents progrès technologiques permettent aujourd'hui d'obtenir de tels volumes rapidement.

### III.5 Comparaison des modèles de génétique d'association

L'analyse de la structuration du panel a révélé l'existence de sous-populations ainsi que d'un apparentement étroit entre paires d'individus, ce qui amène à induire un DL entre locus non physiquement liés. L'utilisation de modèles mixtes prenant en compte la structure et l'apparentement a permis de limiter le risque de détection de fausses associations résultant de la présence de ce type de DL. Yu *et al.*, 2006 ont proposé un modèle combinant un effet fixe dû à la structure et un effet aléatoire qui permet de modéliser la part du phénotype expliqué par l'apparentement entre individus. Ce modèle peut se formuler de la façon suivante :

$$G_i^{BLUP} = \sum_k S_{ik} a_k + M_{il} \theta_l + u_i + e_i$$

$G_i^{BLUP}$  est le BLUP de l'individu  $i$ ,  $S_{ik}$  est la probabilité d'appartenance du génotype  $i$  au groupe  $k$ ,  $a_k$ , l'effet du groupe  $k$  considéré comme fixe,  $M_{il}$  est le génotype de l'individu  $i$  au locus  $l$ ,  $\theta_l$  est l'effet du locus  $l$ .  $u_i$  est l'effet aléatoire modélisant l'apparentement avec  $Var(u) = \sigma_u^2 K_{ais}$  où  $K_{ais}$  est la matrice AIS d'apparentement et  $Var(e) = \sigma_e^2$

Plusieurs méthodes de prise en compte de la structure ont été développées dans ce chapitre ; cependant elles ne couvrent pas toutes le même aspect de la structure, et il est donc nécessaire de comparer ces modèles afin de choisir celui d'entre eux qui permettra de réduire le taux de faux positifs tout en maintenant une puissance suffisante. De plus, il est attendu que l'intérêt de chaque modèle dépende du phénotype. Ainsi, Aranzana *et al.*, (2005) ont montré que des caractères très corrélés à la structure, telle que la floraison dans leur panel d'*Arabidopsis*, entraînaient des taux de faux positifs élevés. L'objectif de cette section est de comparer pour chaque caractère plusieurs modèles statistiques sélectionnés à partir des estimations pertinentes de la structure afin de choisir les meilleurs modèles à appliquer en génétique d'association sur l'ensemble des données génotypiques.

Environnement	Testeur pour lignées B	Testeur pour lignées R
AI08_I	83HR4gms	FS71501
AI08_NI		
CO09_I		
CO09_NI		
GA09_I		
GA09_NI		
LO10		
VE10		
AI09_I	SOLR001M	AT0521
AI09_NI		
VE09_I		
VE09_NI		
CA10		
CO08_I		
CO08_NI		
SE10		
CHA10		

**Table III.4 : Rappel du dispositif d'expérimentation phénotypique**

Modèle		Description		Covariate specification
Naïf		Pas de correction de la structure		-
Effets fixes	Q <sub>2</sub> _ssr	Structure inférée avec le logiciel STRUCTURE	2 groupes 48 SSR	Proportion du génome assigné à chaque groupe
	Q <sub>3</sub> _ssr		3 groupes 48 SRR	
	Q <sub>3</sub> _snp		3 groupes 5923 SNP	
	PC <sub>3</sub> _ssr	Structure résultant de l'ACP	3 composantes principales 48 SSR	Coordonnées sur les composantes principales
	PC <sub>3</sub> _snp		3 composantes principales 5679 SNP	
	Testeur	Structure des pools de lignées (B-pool, R-pool)		Binaire (0/1)
Effets aléatoires	Kais	Apparement estimé à partir de l'AIS		Matrice de Variance-Covariance proportionnelle à l'AIS
Mixtes	Kais+ Q <sub>2</sub> _ssr	Modèles mixtes avec structure en effet fixe et kinship en effet aléatoire		
	Kais+ Q <sub>3</sub> _ssr			
	Kais+ Q <sub>3</sub> _snp			
	Kais+ PC <sub>3</sub> _ssr			
	Kais+ PC <sub>3</sub> _snp			
	Kais+ Testeur			

**Table III.5 : Récapitulatif des modèles testés**



A partir du modèle de Yu *et al.*, (2006) décrit ci-dessus, nous avons comparé plusieurs estimations de la structure obtenues avec le logiciel STRUCTURE ou l'ACP à partir des SNP ou SSR. Un effet du pool d'appartenance des lignées (B ou R) a été ajouté aux comparaisons, étant donné que cette distinction B-R constitue un premier niveau de structure du panel. Cet effet est appelé « effet Testeur » car chaque lignée B ou R est croisée avec un testeur différent selon les environnements (Table III.4). La matrice Q de probabilité d'appartenance à chaque groupe issue du logiciel STRUCTURE, la matrice PC correspondant aux coordonnées sur les composantes principales sélectionnées par Tracy-Widom ainsi qu'un vecteur 0-1 pour l'effet testeur, ont été chacun combiné ou non à une matrice de kinship estimée par le coefficient AIS (Kais). Au total, 14 modèles, synthétisés dans la table III.5, ont été comparés sur chacun des 183 BLUP des caractères phénotypiques.

La première étape a consisté à comparer l'ajustement de chacun des modèles au phénotype, en utilisant le critère BIC défini comme suit :

$$BIC = -2 * \ln(L) + k \ln(n)$$

$\ln(L)$  est le logarithme du maximum de vraisemblance,  $k$  est le nombre de paramètres estimables et  $n$  la taille de l'échantillon.

Les modèles mixtes ont été estimés à partir du maximum de vraisemblance non restreint disponible dans la package R « kinship » (fonction `lmekin`), car ils comportent différents effets fixes.

En complément du BIC, les p-values des effets fixes ont été extraites. La part de variation phénotypique expliquée par chaque modèle a été calculée en utilisant un coefficient de détermination basé sur le rapport de vraisemblance entre le modèle testé et un modèle simplifié avec un seul effet de la moyenne. Cette statistique est formulée dans Sun *et al.*, (2010) de la manière suivante :

$$R_{LR}^2 = 1 - \left(\frac{L_0}{L_M}\right)^{\frac{2}{n}}$$

$L_M$  est le maximum de vraisemblance du modèle d'intérêt,  $L_0$  est le maximum de vraisemblance d'un modèle ne contenant qu'un effet de la moyenne et  $n$ , le nombre d'observations. Ce calcul est adapté aux modèles mixtes et est équivalent au  $R^2$  dans les modèles ne contenant pas d'autres effets aléatoires que l'erreur résiduelle gaussienne.

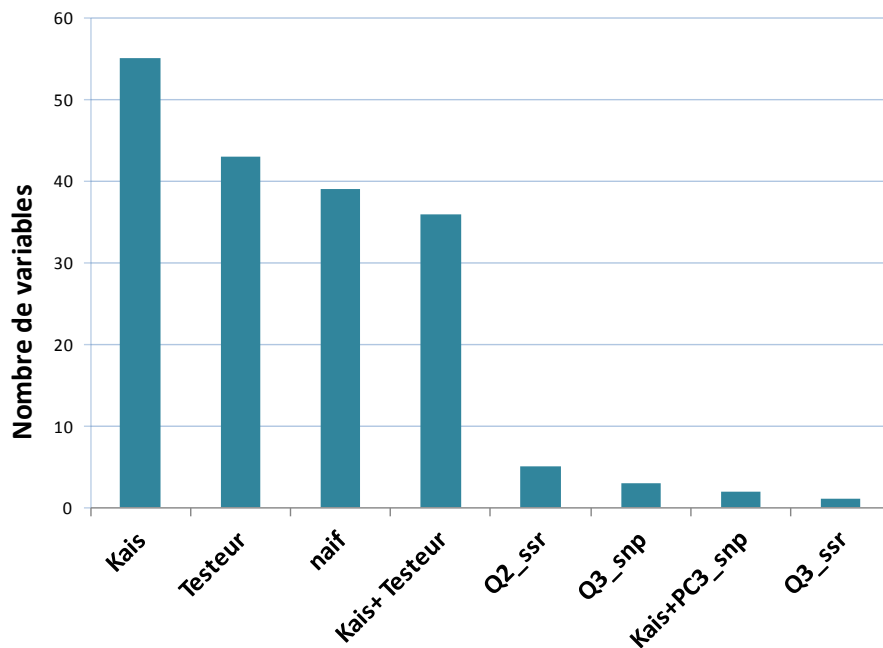


Figure III.12 : Distribution des meilleurs modèles selon le critère BIC sur l'ensemble des variables phénotypiques.

	K <sub>ais</sub>	K <sub>ais</sub> + PC <sub>3_snp</sub>	K <sub>ais</sub> + PC <sub>3_ssr</sub>	K <sub>ais</sub> + Q <sub>2_ssr</sub>	K <sub>ais</sub> + Q <sub>3_SNP</sub>	K <sub>ais</sub> + Q <sub>3_SSR</sub>	K <sub>ais</sub> + Testeur	PC <sub>3_snp</sub>	PC <sub>3_ssr</sub>	Q <sub>2_ssr</sub>	Q <sub>3_snp</sub>	Q <sub>3_ssr</sub>	Testeur
K <sub>ais</sub>													
K <sub>ais</sub> + PC <sub>3_snp</sub>	11.97												
K <sub>ais</sub> + PC <sub>3_ssr</sub>	12.39	3.28											
K <sub>ais</sub> + Q <sub>2_ssr</sub>	4.47	7.93	8.09										
K <sub>ais</sub> + Q <sub>3_SNP</sub>	7.79	4.77	5.21	4.06									
K <sub>ais</sub> + Q <sub>3_SSR</sub>	8.26	4.59	4.74	4.18	1.45								
K <sub>ais</sub> + Testeur	7.79	14.98	15.43	8.27	10.75	11.33							
PC <sub>3_snp</sub>	18.54	12.79	14.64	15.50	12.73	13.30	21.50						
PC <sub>3_ssr</sub>	35.06	25.53	25.14	30.72	28.33	27.75	38.12	17.83					
Q <sub>2_ssr</sub>	25.49	22.72	22.94	23.22	23.17	22.85	30.28	13.24	10.40				
Q <sub>3_snp</sub>	17.08	15.14	16.40	14.95	13.87	14.62	19.92	4.78	19.75	12.54			
Q <sub>3_ssr</sub>	18.93	15.28	16.45	15.93	14.74	14.92	22.53	4.57	16.96	10.14	3.11		
Testeur	17.83	20.64	21.56	18.46	18.42	19.05	14.54	13.06	28.13	19.76	10.20	12.72	
naïf	25.26	23.98	24.14	24.19	24.36	24.08	30.67	14.96	12.03	4.14	14.11	12.01	19.59

Table III.6 : Moyenne des différences de BIC sur l'ensemble des variables entre chaque paire de modèle

La seconde étape de comparaison de modèles réintroduit l'information aux marqueurs, en particulier pour un set de 1000 SNP choisis aléatoirement en dehors des gènes candidats. La capacité de chaque modèle à corriger le taux de faux positifs a été examinée à l'aide de quantile-quantile (QQ plot) en suivant la procédure de Yu *et al.*, (2006). Elle consiste à représenter la distribution des p-values observées en fonction des p-values attendues sous l'hypothèse nulle de non-association entre marqueur et phénotype. Si le modèle corrige correctement de la structure, les points doivent suivre la diagonale car l'on postule qu'il n'y a que rarement d'associations entre les 1000 SNP tirés au hasard et le phénotype. Les fréquences cumulées des p-values sont représentées grâce à la fonction `plotCsum` du package R `PBSmodelling`.

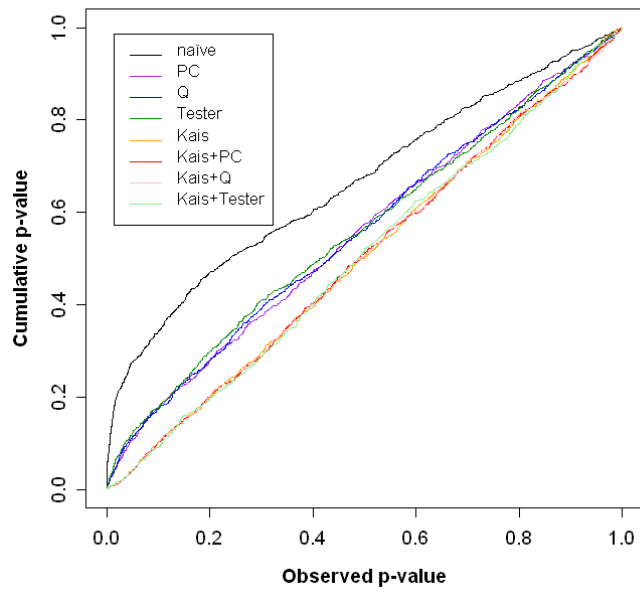
### III.5.2 Résultats et discussion

Le meilleur modèle selon le critère BIC varie d'une variable à l'autre. Cependant, comme indiqué dans la figure III.12, la plupart des variables ont pour meilleur modèle Kais, Testeur, Kais + Testeur ou le modèle naïf. Les résultats basés sur  $R_{LR}^2$  sont instructifs. Par exemple, toutes les variables ayant le modèle naïf comme meilleur modèle n'ont pas de corrélations supérieures à 0.10 avec les covariables de structure, ce qui signifie que la prise en compte des covariables de structure n'est pas nécessaire. Aucune variable n'est corrélée aux covariables Q2\_ssr et PC3\_ssr qui ne ressortent dans aucun « meilleur modèle ».

Des tendances fortes se dégagent selon les lieux. Ainsi Ver09, Ver10, Chateau10, Loudun10 et Aig09 présentent pour la plupart des variables, des corrélations fortes avec les covariables de structure, ce qui n'est pas le cas pour les autres lieux. Les 2 meilleurs modèles associés le plus fréquemment à ce premier groupe de lieux sont les modèles avec testeurs. La part de variance phénotypique expliquée par les covariables de structure atteint des valeurs maximales de 0.45, ce qui laisse penser qu'un manque de puissance est à attendre des tests d'associations avec ces variables.

Peu de tendances se dégagent selon les caractères. Une des seules différences observées concerne les composantes du rendement pour lesquelles des modèles fixes autres que le modèle « Testeur » apparaissent en meilleur modèle.

Les valeurs de BIC étant parfois très proches entre plusieurs modèles, nous avons moyenné les différences entre chaque paire de modèles sur l'ensemble des variables. D'après la table III.6, les modèle naïf, Q2\_ssr et PC3\_ssr, sont très proches l'un de l'autre mais éloignés de l'ensemble des autres modèles, surtout des modèles avec kinship. Les modèles mixtes apparaissent tous très proches. Une fois la kinship spécifié dans un modèle, peu importe la



**Figure III.13 : QQplot représentatif de la plupart des variables**

structure ajoutée, les modèles semblent équivalents. Cependant, le modèle Kais + Testeur apparaît se distinguer. Ce résultat est cohérent avec les p-values observées pour les effets fixes de structure qui ne sont plus significatifs une fois la kinship spécifiée, sauf pour le modèle Kais+ Testeur. La matrice de kinship est donc capable de décrire à la fois les relations d'apparentements qui existent entre paires d'individus mais aussi des niveaux plus globaux de structuration, captés par STRUCTURE et l'ACP. Cependant l'effet testeur est moins bien pris en compte par la kinship.

Quant aux modèles fixes, on observe plus de variabilité entre ces modèles, sans doute révélée par l'absence de kinship. Si l'on compare les modèles de structure inférés à partir des SNP et ceux à partir des SSR, les résultats sont plus proches pour les modèles utilisant les coordonnées de l'ACP que ceux utilisant les probabilités de STRUCTURE. Quant à la comparaison entre modèles avec matrice Q et modèle avec matrices PC, ils sont plus proches dans le cadre d'une inférence avec SSR qu'avec SNP.

La plupart des QQplot, réalisés à partir d'un set de 1000 SNP, rendent compte de la même tendance : seuls les modèles avec kinship sont à même de corriger efficacement l'erreur de type I (exemple Figure III.13 pour la floraison). Nous garderons donc pour les tests d'association le modèle Kais ainsi que, en comparaison, le modèle Kais + Testeur, qui nous permet d'ajouter l'effet très bien contrôlé de la structuration B/R.

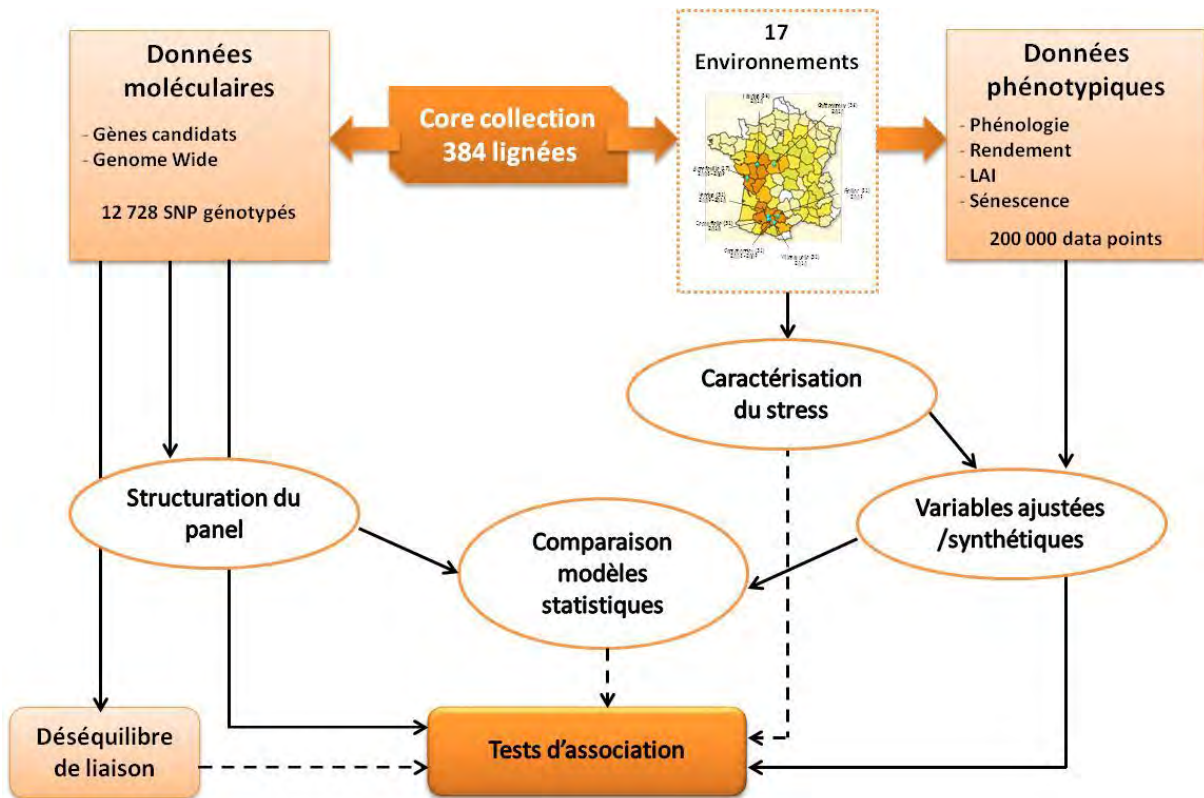


Figure IV.1 Schéma résumant la procédure utilisée pour la recherche de marqueurs associés aux caractères phénotypiques.

# Chapitre IV Locus impliqués dans le déterminisme génétique du rendement sous contraintes hydriques

## IV.1 Introduction

L'objectif de ce chapitre est de décrire les locus identifiés comme impliqués dans la variation génétique du rendement et de ses composantes sous contrainte hydrique. La figure IV.1 résume les différentes étapes permettant d'y aboutir.

La core collection de 384 lignées a été évaluée dans un réseau expérimental de 17 environnements pour une dizaine de caractères liés à la phénologie, la productivité, le maintien de la capacité photosynthétique. Dans le chapitre II, les variables phénotypiques ont été ajustées des hétérogénéités intra environnement à l'aide du meilleur modèle statistique parmi trois modèles testés. Au final, les prédictions génétiques, sous forme de BLUP, de 189 variables phénotypiques ont été extraites pour la génétique d'association. Les stress hydrique et thermique ont été caractérisés dans 14 environnements en utilisant le modèle de culture SUNFLO. Les indicateurs de stress issus de ces analyses ont permis de classer les environnements selon la quantité de stress subi et aussi de construire des variables synthétiques, à partir de régressions linéaires de la productivité en fonction du stress dont les caractéristiques (pente pour la réponse au stress et potentiel à niveau de stress moyen) sont soumises également à l'analyse des associations.

Dans le chapitre III, l'analyse du panel d'association a permis d'identifier une structuration en deux pools de lignées formés par l'histoire de la sélection hybride chez le tournesol: les lignées B, maintenant la stérilité mâle cytoplasmique et les lignées R, restauratrices de la fertilité mâle. La comparaison de modèles statistiques visant à expliquer le phénotype en fonction des différentes matrices de structure et d'apparentement, a abouti à la sélection de deux modèles pour les tests de génétique d'association : un modèle avec une matrice d'apparentement entre toutes les paires d'individus (modèle Kais) et un second modèle avec la matrice d'apparentement et un effet dû au type de lignée B ou R (modèle Kais + Testeur). Ce dernier modèle est formulé de la manière suivante (le modèle Kais exclut les covariables encadrées en pointillées) :





$$G_i^{BLUP} = \sum_k S_{ik} a_k + M_{il} \theta_l + u_i + \varepsilon_i$$

$G_i^{BLUP}$  est le BLUP de l'individu  $i$  pour une combinaison variable - environnement,  $S_{ik}$  indique l'appartenance du génotype  $i$  au groupe  $k$  (pool de lignées B ou R),  $a_k$ , l'effet du groupe  $k$  considéré comme fixe,  $M_{il}$  est le génotype de l'individu  $i$  au locus  $l$ ,  $\theta_l$  est l'effet du locus  $l$ ,  $u_i$  est l'effet aléatoire de l'apparentement entre individus avec  $Var(u) = \sigma_u^2 K_{ais}$  où  $K_{ais}$  est une matrice comprenant les probabilités d'identité par état (AIS : Alike in state) à un locus donné et  $Var(\varepsilon) = \sigma_\varepsilon^2$ .

Les données moléculaires disponibles ont été également décrites dans le chapitre III. Un total de 12 728 SNP provenant pour la plupart d'une approche de détection de SNP non ciblée et pour quelque uns d'une approche gènes candidats, a été génotypé sur la core collection en utilisant différentes plateformes technologiques.

De nombreuses questions se posent à l'orée de ce chapitre. Tout d'abord, étant donné la forte structuration présente dans le panel (Chapitre III), on peut se demander si nous disposerons de suffisamment de puissance pour détecter des marqueurs associés à un phénotype. La présence d'une interaction génotype - environnement a été mise en évidence au cours des chapitres précédents. Cette interaction sera-t-elle un obstacle à la détection de signaux qui persistent sur plusieurs environnements ? Pour prendre en compte cette interaction génotype - environnement, nous avons mis en place plusieurs index multilocaux de réponse au stress, mais seront-ils suffisamment puissants pour détecter des associations ?

Enfin, la complexité du dispositif expérimental constitue un réel défi. Comment interpréter des résultats provenant de testeurs différents selon les lignées et les environnements ?

Dans un premier temps, nous allons exposer les résultats des tests d'association par environnement et par caractère, puis sur les index multilocaux. Enfin nous discuterons de ces différentes méthodes, des locus intéressants et de leurs gènes candidats sous-jacents.

## IV.2 Matériels et méthodes

6628 SNP ayant un taux de données manquantes inférieur à 10% et une MAF supérieure à 5% ont été sélectionnés parmi les 12 728 SNP génotypés. Comme décrit dans le chapitre III, ces



SNP sont issus d'origines diverses : certains ont été ciblés dans des gènes candidats mais la majorité d'entre eux a été défini aléatoirement sur le génome. Par rapport au panel décrit dans le chapitre III, pour l'étude d'association, trente individus supplémentaires ont pu être réintégrés au panel après vérification de leur conformité. Le nombre de lignées à tester est donc passé de 304 à 334. La matrice de Kinship a donc été à nouveau calculée pour ce nouveau set mais la structure n'a pas été réévaluée, considérant que ces quelques lignées supplémentaires ne remettaient pas en cause la majeure partition du panel en lignées B et R.

Tous les tests d'association ont été réalisés dans R avec la fonction `emma.REML.t` (t-test) du package EMMA (Efficient Mixed Model Association), <http://mouse.cs.ucla.edu/emma/news.html>. L'utilisation des modèles mixtes est désormais la méthode de choix pour les études d'association car ils permettent de prendre en compte différents types d'effets confondants. Un des défis important lié à ce type de modèle est la réduction des temps de calculs notamment avec l'accroissement continu de la taille des jeux de données. L'algorithme présent dans EMMA et développé par Kang *et al.*, (2008) est optimisé pour cet objectif. Cependant, il existe aujourd'hui d'autres méthodes (revues dans Zhou *et al.*, 2012) plus sophistiquées et rapides, dont certaines construites à partir de l'algorithme d'EMMA. Kang *et al.*, 2010 ont proposé une méthode d'approximation (EMMAX : EMMA eXpedited) qui évite d'estimer les composantes de la variance à chaque SNP. Zhou *et al.*, (2012) ont développé GEMMA (Genome Wide EMMA) qui, selon les auteurs, a l'avantage d'être une méthode exacte et donc ne risquant pas de diminuer la puissance des tests comme dans les méthodes d'approximation (EMMA appartient également aux méthodes exactes et suffit à la taille de nos matrices d'entrée).

La multiplicité des tests augmente la probabilité de faux positifs, c'est-à-dire de SNP déclarés associés à un phénotype alors qu'ils ne le sont pas. Afin de contrôler le taux de faux positifs, nous avons utilisé deux méthodes classiques: la correction de Bonferroni et le FDR (False Discovery Rate). La correction de Bonferroni consiste à diviser l'erreur de première espèce (exemple 5%) par le nombre de tests indépendants de manière à obtenir un seuil de significativité global. Le nombre de marqueurs indépendants ( $M_{\text{eff}}$ ) a été déterminé grâce au package simpleM [http://sourceforge.net/projects/simplem/files/simpleM\\_Ex.zip/download](http://sourceforge.net/projects/simplem/files/simpleM_Ex.zip/download) qui consiste à éliminer grâce au calcul du DL, les marqueurs corrélés par une méthode de réduction dimensionnelle (Gao *et al.*, 2010). Il nécessite par contre une matrice sans données manquantes. Pour cela, nous avons affecté l'allèle majoritaire au marqueur à chaque allèle manquant.



Le FDR proposé par Benjamin et Hochberg (1995), estime la fraction de faux positifs parmi les SNP déclarés significatifs. Il a été calculé pour chaque test à partir du package R qvalue (<http://www.bioconductor.org/packages/release/bioc/html/qvalue.html>).

Les p-values des tests d'association et les proportions de variances expliquées par les marqueurs ( $R_{LR}^2$ , variance basée sur un ratio de vraisemblance) ont été calculés pour chaque SNP. Les effets alléliques ont été calculés dans ASREML.

Les marqueurs significatifs ont ensuite été localisés sur une carte génétique (privée Biogemma) de 8417 marqueurs dont 8001 sont des SNPs. Cette carte est une carte consensus réalisée à partir de 4 populations de RILs dont la population INEDI, servant de carte initiale, (et qui sera utilisée pour la détection de QTL) et trois populations F2. En ne considérant que les positions différentes sur la carte, les marqueurs sont cartographiés avec une densité moyenne de un tous les 0.57 cM. Les marqueurs n'ayant pas pu être positionnés sur cette carte ont été localisés grâce au calcul de leur déséquilibre de liaison ( $r_{VS}^2$ ) avec l'ensemble des marqueurs cartographiés. En considérant un DL minimum de 0.10, chaque marqueur non présent sur la carte a été localisé à la même position que le marqueur cartographié ayant le DL le plus fort.

### **IV.3 Résultats et discussion**

#### *IV.3.1 Résultats des tests d'association sur chaque combinaison environnement-caractère*

##### *IV.3.1.1 Test multiples*

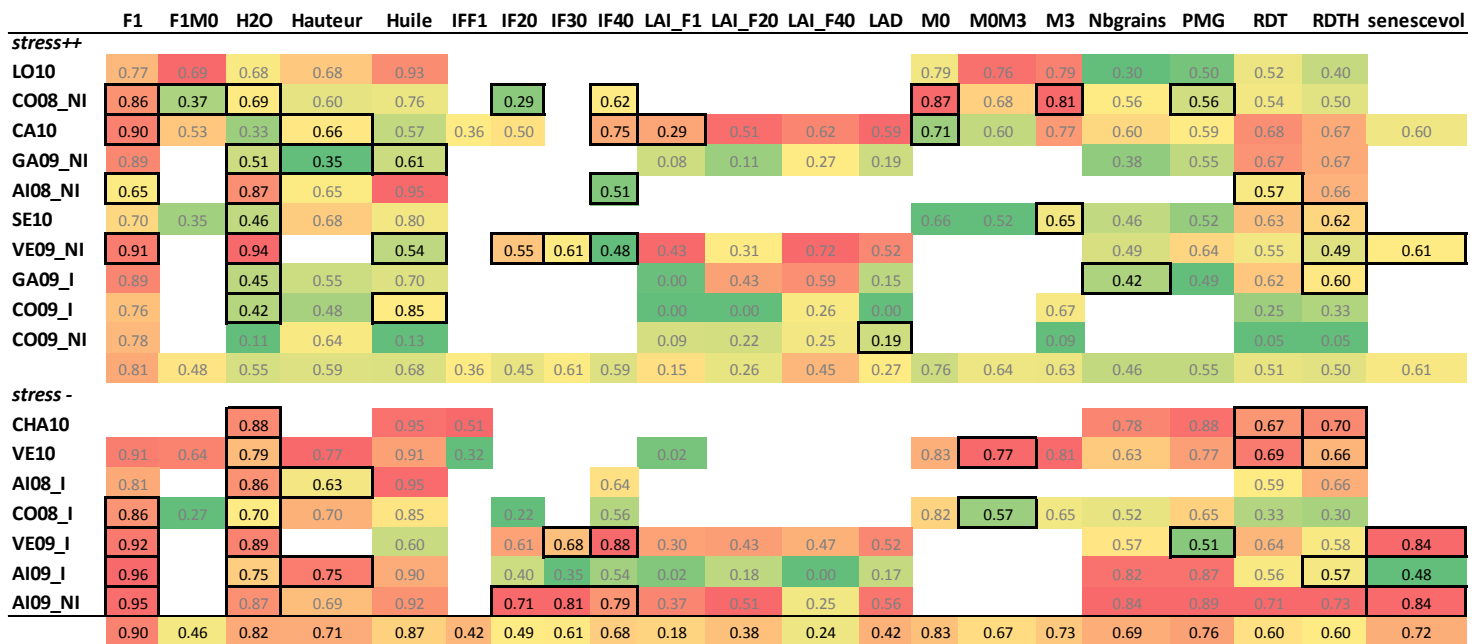
6628 SNP ont été testés indépendamment pour leur association avec 189 variables phénotypiques (combinaisons d'un caractère et d'un environnement) en utilisant 2 modèles statistiques : Kais et Kais + Testeur. Avec un FDR de 10%, 290 SNP présentent au moins une association significative avec un caractère lorsque le modèle choisi est Kais et 157 lorsque le modèle est Kais + testeur. Sur les 6628 SNP testés, 5786 ont été identifiés comme indépendants par le package simpleM, en utilisant l'ensemble des composantes principales expliquant 99.5 % de la variabilité du DL entre marqueurs. Pour une erreur de 5%, le seuil de Bonferroni est donc de  $(0.05/5786) 8.64 \cdot 10^{-06}$ . En utilisant ce seuil, 91 SNP présentent au moins une association significative avec le modèle Kais et 55 avec le modèle Kais + Testeur. Ces résultats soulignent que même si le nombre de SNP significatifs diminue avec la correction de Bonferroni, il reste cependant un nombre non négligeable de SNP avec des p-values très faibles (minimum à  $1.99 \cdot 10^{-08}$ ), y compris dans certains cas pour des caractères



complexes tels que le rendement. Cependant, l'utilisation d'une correction trop conservatrice sacrifie une partie de la puissance des tests (Muller *et al.*, 2010). Au contraire le FDR, est connu pour être plus souple mais il nécessite que les marqueurs ne soient pas considérablement en DL (Chen et Storey, 2006), ce qui semble être le cas étant donné que seul 87 % des marqueurs testés ont été estimés indépendants avec le package simpleM. Le FDR représente la proportion de faux positifs parmi les marqueurs significatifs obtenus. Il a été également montré par Chen et Storey (2006) que cette méthode était bien adaptée à l'analyse de nombreux caractères simultanément (dans notre cas, 189 variables). Après avoir fait une ACP sur la matrice des 189 BLUP, 40 composantes principales permettent d'expliquer 80% de la variabilité phénotypique. Contrairement au FDR pour lequel la proportion de faux positif serait la même sur l'ensemble des caractères, la correction de Bonferroni exigerait de diminuer le risque de chaque test en divisant à nouveau par 40, pour conserver un risque global contrôlé. Nous avons donc fait le choix de discuter à partir des résultats obtenus avec un FDR de 10%, et de confronter d'autres critères aux p-values et aux variances expliquées par les marqueurs ( $R_{LR}^2$ ), tels que les effets alléliques.

#### IV.3.1.2 Comparaison des modèles Kais et Kais + Testeur

La différence entre les deux modèles statistiques d'association (Kais et Kais + Testeur) est claire : sur les 741 associations marqueur-phénotypes significatives, seules 250 (34%) sont communes aux deux modèles. Parmi les associations détectées sur un seul modèle, le modèle majoritaire (88%) est le modèle Kais. Lorsque l'on regarde la position sur la carte génétique de ces marqueurs trouvés associés seulement avec le modèle Kais, la quasi-totalité (84%) est cartographiée sur les chromosomes 10 et 13. Ces chromosomes présentent des grands blocs de déséquilibre de liaison (incluant ces SNP détectés), autour des locus de ramification (pour le LG10) et de restauration de la fertilité (pour le LG13). Ce phénomène, appelé hitchhiking (Maynard *et al.*, 1974), survient lorsqu'un locus sélectionné entraîne l'extension du DL autour de ce locus. Le locus de ramification joue un rôle crucial dans les programmes de sélection chez le tournesol puisque la ramification des lignées R conduit à une plus large fenêtre d'hybridation avec les lignées femelles dans les champs de production de semences hybrides (Bachlava *et al.*, 2010). Le LG13 est lui caractérisé par la présence d'un locus essentiel car il restaure la fertilité mâle des lignées CMS (cytoplasmic male sterile). La sélection des phénotypes ramifiés et restaureurs de fertilité dans les lignées R a probablement entraîné un niveau de structuration supplémentaire par rapport à celui dont rend compte l'apparentement. Il est donc vraisemblable que l'introduction de l'effet B/R dans le modèle a



**Figure IV.2 : Héritabilité par caractère et environnement.** L'échelle de couleur du vert au rouge est proportionnelle à l'héritabilité définie variable par variable. Les valeurs encadrées correspondent aux variables pour lesquelles des associations ont été détectées avec le modèle Kais + Testeur et un FDR de 10%. Les environnements sont classés selon le stress défini par le modèle SUNFLO (sauf CO09\_NI, AI09\_I et AI09\_NI, mal prédits par SUNFLO).



permis d'éliminer les SNP faux positifs associés à cette couche supplémentaire de structuration.

Concernant les autres chromosomes, le modèle Kais ne présente pas d'inflation majeure dans le nombre de SNP détectés. On note par ailleurs que le modèle Kais + Testeur permet, sur ces autres chromosomes, d'identifier des associations supplémentaires, notamment pour le rendement. Il est donc possible que la réduction des effets confondants ait augmenté la puissance de détection car ceux-ci avaient tendance à masquer le vrai signal (Zhao *et al.*, 2007).

Nous nous intéresserons principalement aux résultats du modèle Kais + Testeur, qui permettra de sélectionner des SNP d'intérêt indépendamment du type de lignée (B ou R). Les SNP détectés uniquement avec le modèle Kais en dehors des blocs de DL sur le LG10 et LG13 sont rares mais pourront tout de même être observés, car parmi eux existent peut être des vrais positifs.

#### *IV.3.1.3 Résultats du modèle Kais + Testeur selon les environnements*

La figure IV-2 représente la répartition des associations significatives détectées par environnement et par caractère (encadrés), et les valeurs d'héritabilités issues du modèle naïf selon un gradient de couleur. Les environnements sont classés selon le stress estimé par le modèle SUNFLO.

Des associations ont pu être détectées sur la plupart des environnements. Parmi les lieux où très peu d'associations ont été détectées, LO10 est typé comme « stressant » par le modèle SUNFLO avec un nombre de jours de stress cumulant à plus de 47 jours autour de la floraison, quel que soit le génotype témoin considéré. De même, les autres environnements stressants, (AI08\_NI, CA10, CO09\_I, GA09\_I, GA09\_NI et SE10), présentent peu d'associations. Ces 7 environnements ont en moyenne une héritabilité plus faible que sur les lieux non stressants, parmi lesquels, VE09\_I, AI09\_I, CHA10 et VE10 sont marqués par une variance génétique plus forte et un plus grand nombre d'associations détectées.

La diminution de l'héritabilité avec un stress hydrique croissant a été très souvent mentionnée dans la littérature pour expliquer la difficulté d'obtenir un progrès génétique important dans ces conditions, notamment pour un caractère complexe tel que le rendement (Frahm *et al.*, 2004 ; Tuberosa *et al.*, 2002; Messmer *et al.*, 2009).

Cette héritabilité plus faible provient de la variabilité spatiale, plus forte en conditions de stress hydrique par exemple du fait d'une hétérogénéité dans la profondeur du sol, qui augmente la variance environnementale. La relation entre rendement et apports en eau n'étant pas linéaire (Blum, 2011), lorsque les apports sont à un niveau très faible, la moindre



variation spatiale a un impact important sur le rendement. Ceci souligne l'intérêt d'améliorer le dispositif au champ et de mieux contrôler cette variabilité afin d'augmenter l'héritabilité. De plus, nous pouvons également nous demander si le panel d'association utilisé dans cette étude possède suffisamment de variabilité pour la réponse au stress hydrique. En effet, plusieurs études ont obtenu une héritabilité plus forte sous conditions hydriques limitantes en utilisant des populations issues de parents présentant une variabilité pour la tolérance au stress hydrique (Ceccarelli *et al.*, 1998, Venuprasad *et al.*, 2007).

Le seul caractère ayant une héritabilité plus forte dans le groupe de lieux stressés est l'indice foliaire à 40 jours (IF40). Ce caractère, qui pourrait donc être un caractère d'adaptation à la sécheresse révèle un intérêt majeur. Le « stay green » qu'il traduit est donc plus discriminant en conditions hydriques limitantes (Cairns *et al.*, 2012). Plusieurs associations ont ainsi été détectées sur des environnements stressants comme CO08\_NI, CA10, VE09\_NI et AI08\_NI.

Cependant l'héritabilité moyenne plus faible pour le groupe des « environnements stressés » n'est que partiellement explicative du plus faible nombre d'associations détectées. En effet, ces environnements ont autant, voire plus de caractères pour lesquels au moins une association a été détectée. A l'inverse, des environnements à potentiel tels que AI09\_NI ne permettent pas la détection de pics d'association sur le rendement et ses composantes pour lesquels l'héritabilité est forte. La figure IV-2 montre que les associations détectées ne correspondent donc pas forcément aux variables ayant les plus fortes héritabilités.

#### *IV.3.1.4 Résultats du modèle Kais + Testeur selon les caractères*

Des associations ont été détectées pour tous les caractères à part l'indice foliaire et le LAI à floraison qui ne permettent pas de bien différencier les génotypes (cf. chapitre II). Par contre les indices foliaires mesurés à des stades plus tardifs présentent des associations dans la plupart des environnements où ils ont été mesurés. Leur héritabilité est meilleure que celle du LAI.

Les stades phénologiques permettent d'identifier des associations dans quasiment tous les environnements, l'héritabilité étant également en général assez forte pour ces caractères. Par contre les variables de productivité ne ressortent que sur quelques lieux. Le rendement en huile apparaît le plus fréquemment (sur 6 environnements) alors que le rendement en grain, qui est pourtant très corrélé au rendement en huile, n'a de SNP associés que sur 3 environnements (différents de ceux du rendement en huile). Le déterminisme génétique pour le rendement et la teneur en huile pourrait donc être différent.



Peu d'environnements permettent de détecter des associations avec les composantes du rendement: deux pour le poids de 1000 grains et un seul pour le nombre de grains. Dans notre dispositif expérimental, ces caractères élémentaires se révèlent donc moins intéressants que prévu (rôle de ces variables pour décomposer le rendement), notamment dans les lieux stressés où leur héritabilité reste faible, alors que dans ces mêmes lieux l'héritabilité du rendement peut atteindre des valeurs aussi élevées que dans les lieux non stressés.

Enfin, la teneur en huile, caractère figurant parmi les plus héréditaires, présentent très peu d'associations.

L'héritabilité moyenne des caractères et des environnements n'est donc pas totalement corrélée au nombre d'association détectées. A cela, plusieurs explications sont possibles. Tout d'abord, il peut s'agir de caractères au déterminisme génétique oligogénique où nous aurions des effets forts à chaque locus mais non détectés faute de marqueurs dans ces régions. La suite de ce chapitre démontrera que le déterminisme est plutôt polygénique. Il peut également s'agir de problèmes de manque de puissance. Les allèles rares sont ainsi indétectables par cette approche de génétique d'association. La densité de marquage disponible dans cette étude n'est pas non plus optimale. Etant donnée l'étendue du déséquilibre de liaison, il faudrait en moyenne deux fois plus de SNP (soit environ 12 000) pour disposer d'une bonne couverture du génome. De plus, la structure du panel élimine automatiquement les marqueurs dont les fréquences alléliques sont corrélées à cette structure, même dans le cas où ils sont vraiment associés au caractère.

Un autre facteur essentiel dans cette étude est lié au dispositif expérimental. L'observation de la performance des lignées à travers leur combinaison avec plusieurs testeurs a un effet sans doute très important. L'utilisation de lignées élites comme testeur pourrait avoir masqué l'effet des allèles propres aux lignées. C'est peut-être ce que le caractère « teneur en huile » illustre. En effet, l'erreur environnementale est très faible mais les testeurs utilisés étant performants pour ce caractère, écrasent la variabilité possible. Le choix des testeurs est une question importante. L'utilisation de lignées élites s'expliquent par la nécessité d'écartier les problèmes de maladies et autres défauts agronomiques pouvant biaiser l'évaluation de la productivité. De plus, pour les sélectionneurs impliqués dans ce projet, elle permet d'identifier des combinaisons hybrides intéressantes pour le marché. Mais on peut se demander s'il n'aurait pas fallu choisir des testeurs non élites afin de mieux révéler la variabilité du panel.



#### IV.3.1.5 Zones génomiques détectées

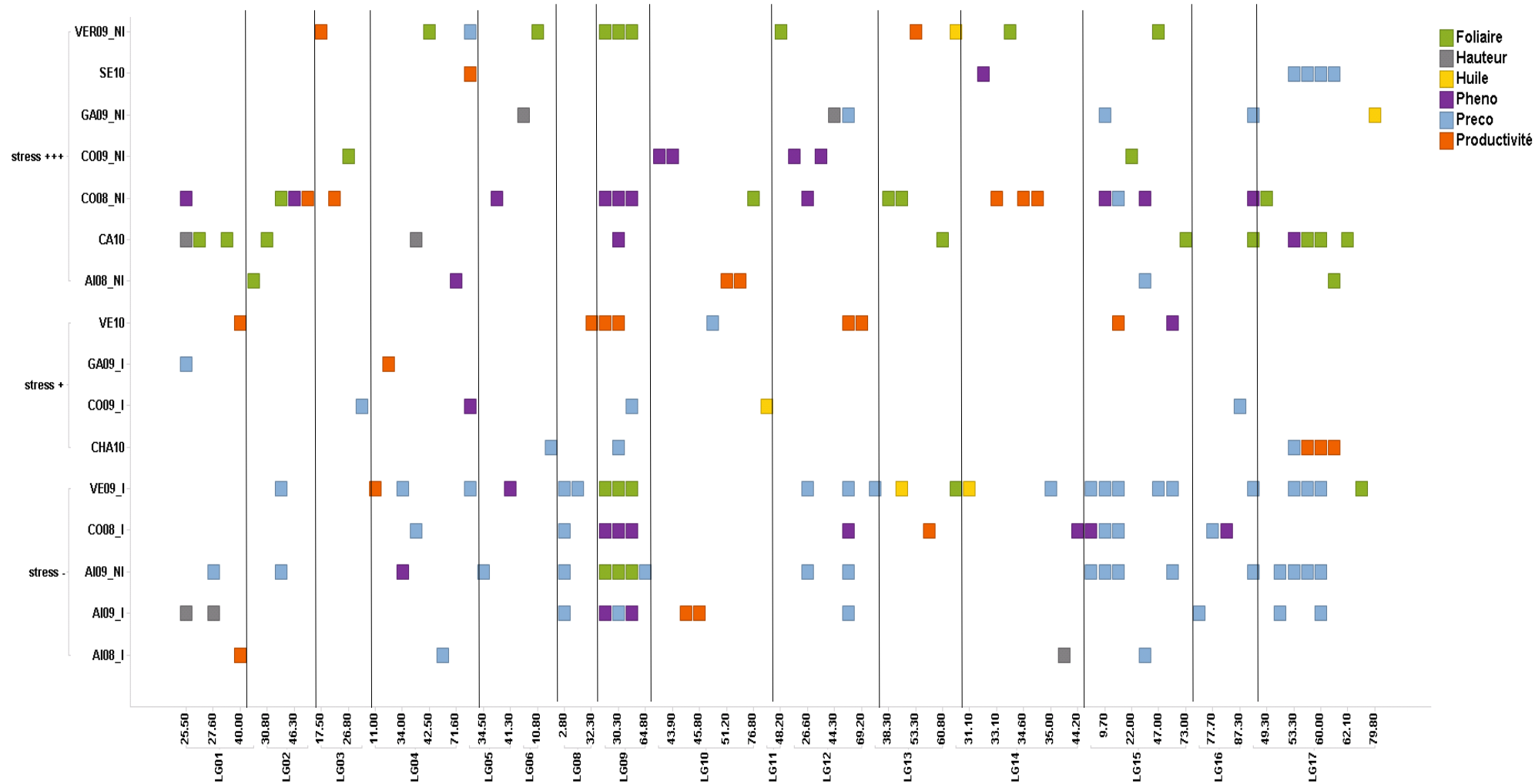
##### - Déterminismes génétiques

Les statistiques (p-values,  $R_{LR}^2$ , effets alléliques pour le modèle Kais + Testeur et un FDR de 10%) sont disponibles en rapport annexe. Les p-values varient entre  $3.38 \cdot 10^{-4}$  à  $1.99 \cdot 10^{-8}$  ( $4.31 \cdot 10^{-5}$  en moyenne), et les  $R_{LR}^2$  entre 4 et 9% mais la médiane se situe entre 0.05 et 0.06. Le  $R_{LR}^2$  varie donc très peu et ne dépasse pas 10%. Ces résultats sont assez classiques dans la littérature, chez les plantes la plupart des études de génétique d'association expliquant entre 5 et 20% de la variabilité phénotypique (Ingvarsson *et al.*, 2010). Les effets très faibles (variance expliquée par un marqueur inférieure à 4%) ne sont pas détectables, peut être en raison du nombre d'individus insuffisant dans ce panel. Quant au seuil maximum de 10%, il révèle à la fois la possibilité d'une héritabilité insuffisante (Brachi *et al.*, 2011), due notamment au fait que les allèles rares ne sont pas pris en compte, et aussi la présence d'un déterminisme génétique complexe avec beaucoup de locus expliquant chacun une faible part de variance phénotypique. Les valeurs de  $R_{LR}^2$  ne varient quasiment pas entre caractères ni entre modèle d'association (Kais ou Kais + Testeur).

Sur les 157 marqueurs associés, 46 proviennent de gènes candidats (soit 3.46 % des gènes candidats initialement testés) et 111 de l'approche non ciblée (soit 2.09% du set initial). D'autre part, un test de chi-deux sur la table de contingence des variables « être ou non associé » et « être ou non dans un gène candidat » rejette l'hypothèse d'indépendance (p-value = 0.003) On peut donc en conclure que la pré-sélection des gènes candidats n'a pas été vaine et qu'elle a amélioré la puissance de détection de l'analyse.

93 SNP ont été cartographiés sur la carte consensus privée parmi les 157 pour lesquels au moins une association avait été détectée. Pour les 64 SNP n'ayant pas de position, le DL ( $r_{VS}^2$ ) a été calculé avec les marqueurs cartographiés ayant été génotypés sur le panel (5511 marqueurs). 55 SNP ont ainsi pu être localisés avec des valeurs de DL supérieures à 0.10.

La figure IV.3 représente la localisation des associations sur le génome par environnement et par type de caractère. Le terme « foliaire » regroupe toutes les variables de LAI et d'indice foliaire. Le terme « phénologie » regroupe les stades (F1, M0, M3) et durée de stades et le terme « productivité » rassemble le rendement en grain et en huile, le nombre de grains, le poids de mille grains et la durée M0M3 de remplissage du grain, variable corrélée au rendement et donc associée aux composantes du rendement (cf. Chapitre II). Les autres termes précocité à la récolte (H2O), teneur en huile et hauteur recouvrent un unique caractère.



**Figure IV.3: Localisation des associations sur la carte génétique en fonction des environnements classés par potentiel (stress+++ correspond à JS > 30 jours de stress cumulé, stress + correspond à JS >20 et stress – correspond à JS <10 (AI09\_I, AI09\_NI et CO09\_NI ont été affectés en fonction de leur potentiel).**



Des associations ont été détectées sur tous les chromosomes. Certains caractères montrent un déterminisme génétique très polygénique. Ainsi la précocité est associée avec des zones génomiques sur 15 chromosomes. La floraison, le stade M3 ou encore l'IF40 présentent des associations sur de nombreuses zones différentes. Les variables de productivité révèlent également un déterminisme polygénique, même si le nombre de zones différentes détectées est plus faible (exemple : six zones sur six chromosomes pour RDTH).

Chez le tournesol, plusieurs études ont également mis en évidence le déterminisme polygénique de certains caractères tels que la teneur en huile (Bert *et al.*, 2003, Othmane *et al.*, 2012) et même la floraison (Stoenescu, 1974 ; Machacek *et al.*, 1979). Pourtant, pour le caractère teneur en huile, nous n'avons identifié que trois zones. Ces résultats ne permettent donc pas d'identifier des gènes majeurs et confirment la complexité du déterminisme de plusieurs caractères d'intérêt agronomique chez le tournesol. Ceci risque de contraindre la sélection assistée par marqueurs puisque de nombreux gènes, chacun ayant un faible effet sont responsables des phénotypes observés.

#### - Relation rendement et précocité

D'après la figure IV.3, les locus associés à la précocité à la récolte colocalisent avec des caractères issus des catégories phénologie et foliaire (par exemple sur les LG09, LG15 et LG17). En général, les marqueurs ayant un effet positif sur la teneur en eau (c'est-à-dire synonymes de tardiveté), ont également un effet positif sur l'allongement des phases phénologiques (plus tardifs), sur les indices foliaires (maintien plus long de la photosynthèse) et sur le rendement. Ce résultat est souvent observé dans des milieux non contraints, où une durée de cycle plus longue favorise le rendement.

Pour quelques marqueurs associés au rendement, l'allèle favorable pour le rendement est également associé à la précocité sur d'autres environnements, mais il s'agit de résultats obtenus qu'avec le modèle Kais. C'est le cas de marqueurs situés sur deux gènes candidats positionnés au même endroit sur le LG17. Les allèles favorables pour RDT (+1.09 qtx/ha) et RDTH (+ 0.67 qtx/ha) sur CO09\_I amènent une précocité de floraison de 1.96 jours et une précocité à la récolte (H2O) de 0.71%. Un autre marqueur sur le LG14 possède un allèle favorable pour le rendement en huile sur LO10 (+1.30) ainsi que pour la teneur en huile dans plusieurs environnements et qui a pour autre effet une arrivée au stade M0 plus rapide (1.99 jours) sur SE10.



Ces lieux sont de nature stressante, notamment CO09\_I qui possède un stress précoce d'après les résultats du modèle SUNFLO. Il est donc possible que les génotypes ayant fini leur cycle plus rapidement aient réussi à réaliser leur potentiel de rendement en évitant la sécheresse. Ces deux zones révèlent donc un intérêt en sélection car elles permettraient d'améliorer le rendement en conditions stressantes. Toutefois, avec le modèle Kais + Testeur, toutes ces associations ne sont pas retrouvées, notamment pour le rendement. S'il peut s'agir de faux positifs dus à la structuration du panel en lignées B/R, il pourrait aussi s'agir d'un manque de puissance. Il est donc intéressant d'approfondir ces résultats.

#### - Interaction GE

On observe une concentration importante d'associations dans trois zones génomiques sur les LG09, LG15 et LG17. Les marqueurs identifiés dans ces régions sont associés principalement à des caractères foliaires et de phénologie sur de nombreux environnements différents (jusqu'à 10 environnements pour le LG17). Ces zones présentent des p-values en générales plus faibles que les autres zones, allant jusqu'à  $1.88 \cdot 10^{-8}$  sur le LG09 pour IF40. Au contraire, le plus souvent, les associations relatives à la productivité sont spécifiques d'une catégorie d'environnement (de potentiel à très stressé). Ce résultat traduit l'importance de l'interaction GE pour le caractère intégrateur que représente la productivité. C'est pourquoi, au-delà de cette analyse caractère par caractère et lieu par lieu, nous nous sommes intéressés à prendre en compte la variabilité du stress hydrique exprimée au travers du réseau d'expérimentation.

### *IV.3.2* Résultats et discussion des associations sur différents index de réponse au stress au travers du réseau multilocal

#### *IV.3.2.1* Introduction

La section II.3 décrit la mise en place des index multilocaux. Pour rappel, le modèle SUNFLO nous a permis de caractériser les stress hydrique et thermique présents dans chaque environnement grâce aux simulations réalisées à partir de trois variétés témoins : MELODY, PACIFIC et PEGASOL. Plusieurs indicateurs de stress à chaque stade du cycle de culture, (avant, autour et après floraison) ont été extraits du modèle et comparés pour leur capacité à prédire le rendement des témoins. Des analyses en composantes principales ont permis de montrer que le stade de culture n'avait pas d'influence majeure sur la prédiction du rendement.

<b>Variable de réponse</b>	<b>RDT</b>	<b>RDTH</b>
<b>Covariable (s)</b>	<b>JS</b>	<b>JS et IFT</b>
Ordonnée à l'origine	RDT_moy	RDTH_moy
Coefficient de régression covariable 1	RDT_TOLH	RDTH_TOLH
Coefficient de régression covariable 2	X	RDTH_TOLT
Coordonnées sur la PC2 covariable 1	RDT_SELH	RDTH_SELH
Coordonnées sur la PC2 covariable 2	X	RDTH_SELT

Table IV.1 : Dénomination des variables synthétiques.

De plus, parmi les variables phénotypiques de productivité, le PMG (poids de mille grains) est très mal représenté sur le plan factoriel de l'ACP.

Des régressions multiples visant à expliquer les variables de réponses choisies : RDT (rendement grains), RDTH (rendement en huile), ou Nbgrains (nombre de grains), par les indicateurs de stress : cumul de stress hydrique (JS) et cumul de stress thermique (IFT) tout au long du cycle, ont été ensuite menées. Malgré quelques différences sensibles entre génotypes et entre variables de réponse, le modèle à deux covariables (JS + IFT), redéfini ci-dessous, était celui qui ressortait le plus souvent comme meilleur modèle d'après le critère AIC, pour les trois variables de réponse. Ce modèle a donc été ensuite appliqué aux lignées du panel d'association en considérant la moyenne de stress entre les 3 variétés témoins.

Le modèle de régression suivant a été appliqué pour chaque génotype du panel en utilisant la fonction lm de R.

$$Y_{ij} = a_i + b_i JS_j + c_i IFT_j + \varepsilon_{ij}$$

avec  $Y_{ij}$  le BLUP de la variable de réponse pour la variété  $i$ ,  $a_i$ , la variable de réponse potentielle,  $JS_j$ , la moyenne du nombre de jours de stress hydrique des trois variétés témoins pour l'environnement  $j$ ,  $b_i$ , la pente associée au stress hydrique,  $IFT_j$ , la moyenne du stress thermique des trois variétés témoins pour l'environnement  $j$ ,  $c_i$ , la pente associée au stress thermique,  $\varepsilon_{ij}$ , l'erreur résiduelle.

Pour chaque caractère de productivité (RDT, RDTH), plusieurs variables synthétiques ont été testées en génétique d'association :

- L'ordonnée à l'origine de la régression du caractère sur la covariable de stress (JS ou IFT), qui représente la performance lorsque le stress est moyen
- Le ou les coefficients de régression traduisant la tolérance au stress hydrique ou thermique
- L'indice de sélection combinant la performance potentielle lorsque le stress est moyen et la stabilité : il correspond aux coordonnées de la deuxième composante principale de l'ACP réalisée sur la matrice (ordonnée à l'origine, coefficient de régression)

Toutes les dénominations de ces variables synthétiques sont indiquées table IV.1



#### IV.3.2.2 Résultats des analyses d'associations sur les variables synthétiques

Les tests de détection d'associations entre le set de 6628 marqueurs et les variables synthétiques liés au rendement grains et huile ne révèlent que très peu de SNP significatifs avec le modèle Kais + Testeur. Seul 1 SNP est significatif avec un FDR de 10% pour RDTH\_moy. Les graphes représentant les p-values des marqueurs testés sur l'ensemble des chromosomes (« Manhattan plot ») sont disponible en annexe (Figure S0). Un seuil a été fixé à  $-\log_{10}(p)=3$ , ce qui correspond à des p-values à  $10^{-3}$  permettant de mettre en évidence quelques zones intéressantes. La figure IV.5 résume les principales zones selon leur position génétique et la variable synthétique.

Les variables de rendement en grain et en huile lorsque le stress est moyen (RDT\_moy ou RDTH\_moy) totalisent le plus de signaux, sur un total de 12 chromosomes, RDT\_moy permettant de détecter plus d'associations. Même si certains marqueurs sont associés à la fois à RDTH\_moy et RDT\_moy (sur les LG05 et LG13), d'autres sont spécifiques d'un type de variable (rendement en huile sur les chromosomes 1 à 3 et 16, rendement grains sur les chromosomes 8,12 et 15). Le rendement en huile et le rendement en grains sont des variables très corrélées, cependant leur déterminisme génétique n'est pas identique. Ce résultat confirme donc l'intérêt d'utiliser ces deux variables pour construire des variables synthétiques. Le rendement potentiel n'est cependant pas un des caractères les plus utiles car il ne permet que de prédire le potentiel des génotypes dans des conditions moyennes de stress, mais ce stress est attendu très variable d'un environnement à l'autre.

Les index de tolérance au stress hydrique obtiennent peu de marqueurs significatifs. Pour le rendement en grains, une seule zone est révélée sur le LG03, zone n'étant pas en DL avec les marqueurs associés à d'autres variables synthétiques sur ce chromosome. Pour le rendement en huile, davantage de zones paraissent intéressantes: 4 marqueurs répartis sur 3 chromosomes (LG05, LG10, LG12) présentent des p-values significatives par rapport au seuil fixé. Sur le LG05, l'un de ces marqueurs est d'ailleurs également associé à l'indice de sélection du rendement en grains (RDT\_TOLH).

Les variables synthétiques liées au stress thermique soulignent des régions génomiques situées sur des chromosomes différents de celles liées au stress hydrique (Figure IV.5). Sur le LG14, 11 marqueurs sont associées avec l'index de tolérance au stress thermique (RDTH\_TOLT) ou l'indice de sélection RDTH\_SELT.

LG	SNP	Position	Variable	Modèle	Modèle synthétique	P-values	R2	Effets allèle favorable	Associations univariées
LG01	HS092129	40	RDTH	RDTH =f(JS + IFT)	RDTH_moy	9.30E-04	0.01	0.33	RDTH_AI08_I, RDTH VE10
LG02	HS127806	7.1	RDTH	RDTH =f(JS + IFT)	RDTH_SELH	2.93E-04	0.04	0.25	
	HS061416	17	RDTH	RDTH =f(JS + IFT)	RDTH_moy	7.52E-04	0.03	0.15	
	HS086468	17	RDTH	RDTH =f(JS + IFT)	RDTH_moy	5.03E-04	0.03	0.06	
	HS086468	17	RDTH	RDTH =f(JS + IFT)	RDTH_SELH	7.11E-04	0.04	0.57	
LG03	HS089004	26.8	RDT	RDT=f(JS)	RDT_TOLH	3.45E-04	0.04	2.40	
	HS155622	36.3	RDTH	RDTH =f(JS + IFT)	RDTH_moy	1.90E-04	0.01	0.48	RDTH_VE09_NI (K)
	HS155620	36.3	RDTH	RDTH =f(JS + IFT)	RDTH_moy	2.43E-04	0.01	0.54	RDTH_VE10(K)
	HS125748	44.9	RDT	RDT=f(JS)	RDT_TOLH	5.55E-04	0.02	1.75	
LG05	HS136674	-0.2	RDT	RDT=f(JS)	RDT_moy	4.80E-04	0.02	0.22	
	HS136674	-0.2	RDTH	RDTH =f(JS + IFT)	RDTH_moy	5.95E-04	0.02	0.08	
	HS136674	-0.2	RDTH	RDTH =f(JS + IFT)	RDTH_SELH	9.81E-04	0.04	0.01	
	HS062748	4.2	RDT	RDT=f(JS)	RDT_SELH	7.92E-04	0.04	0.20	
	HS062748	4.2	RDTH	RDTH =f(JS + IFT)	RDTH_TOLH	4.58E-04	0.02	0.68	
	HS067451	4.2	RDTH	RDTH =f(JS + IFT)	RDTH_TOLH	8.61E-04	0.01	0.65	
	HS114036	26	RDT	RDT=f(JS)	RDT_moy	9.12E-04	0.01	0.15	
	HS147113	34.3	RDT	RDT=f(JS)	RDT_moy	3.28E-04	0.02	0.41	H2O CO08_NI, M3 CO09_I
	HS147124	34.3	RDTH	RDTH =f(JS + IFT)	RDTH_SELH	9.31E-04	0.03	0.39	RDTH SE10
	HS098570	42.1	RDTH	RDTH =f(JS + IFT)	RDTH_TOLH	7.69E-04	0.03	0.05	
HS124266	46.6	RDT	RDT=f(JS)	RDT_SELH	8.33E-04	0.04	0.12		
HS071099	47.3	RDTH	RDTH =f(JS + IFT)	RDTH_SELH	7.54E-04	0.04	0.11		
LG06	HS071322	19.9	RDTH	RDTH =f(JS + IFT)	RDTH_SELT	8.96E-04	0.04	0.20	
LG08	HS082099	15.9	RDT	RDT=f(JS)	RDT_moy	9.47E-04	0.03	0.27	
	HS092155	15.9	RDT	RDT=f(JS)	RDT_moy	9.47E-04	0.01	0.27	
LG09	HS106242	30.3	RDT	RDT=f(JS)	RDT_moy	5.07E-04	0.03	0.26	Hauteur, H2O,F1,IF,M0,Senes,RDT*
	HS068263	34.4	RDT	RDT=f(JS)	RDT_SELH	8.15E-04	0.04	0.01	
	HS144900	53.6	RDT	RDT=f(JS)	RDT_SELH	6.39E-04	0.04	0.04	
LG10	HS148756	43.9	RDT	RDT=f(JS)	RDT_moy	9.37E-04	0.05	0.48	F1 CO09_NI
	HS164332	49	RDTH	RDTH =f(JS + IFT)	RDTH_TOLH	5.26E-04	0.02	0.32	
	HS132684	51.7	RDT	RDT=f(JS)	RDT_moy	4.06E-04	0.01	0.20	RDT_AI08_NI
HS148478	65.1	RDT	RDT=f(JS)	RDT_SELH	9.34E-04	0.04	0.08		
LG11	HS123494	43.4	RDTH	RDTH =f(JS + IFT)	RDTH_SELT	9.09E-04	0.04	0.14	
LG12	HS142046	50.3	RDTH	RDTH =f(JS + IFT)	RDTH_TOLH	4.61E-04	0.03	0.57	
	HS084642	56.8	RDT	RDT=f(JS)	RDT_moy	4.33E-04	0.002	0.76	RDT VE10, H2O**,M3**
LG13	HS080048	12.5	RDT	RDT=f(JS)	RDT_moy	1.35E-04	0.03	0.03	
	HS152545	38.9	RDTH	RDTH =f(JS + IFT)	RDTH_moy	4.50E-04	0.01	0.09	RDTH VE09_I NI (K)
	HS070817	38.9	RDTH	RDTH =f(JS + IFT)	RDTH_moy	4.50E-04	0.01	0.09	RDTH VE09_I (K)
	HS082423	46.2	RDT	RDT=f(JS)	RDT_moy	8.04E-04	0.003	0.21	RDTH* senes (K)
	HS082423	46.2	RDTH	RDTH =f(JS + IFT)	RDTH_moy	2.11E-05	0.01	0.10	
HS147157	53.3	RDTH	RDTH =f(JS + IFT)	RDTH_SELH	3.09E-04	0.04	0.21	RDTH VE09_NI (K)	
LG14	HS085794	29.6	RDTH	RDTH =f(JS + IFT)	RDTH_SELT	9.28E-04	0.04	0.10	
	HS093185	29.6	RDTH	RDTH =f(JS + IFT)	RDTH_SELT	1.14E-04	0.05	0.16	
	HS151070	31	RDTH	RDTH =f(JS + IFT)	RDTH_TOLT	1.66E-04	0.05	0.41	
	HS151076	31	RDTH	RDTH =f(JS + IFT)	RDTH_TOLT	1.06E-04	0.03	0.41	
	HS151077	31	RDTH	RDTH =f(JS + IFT)	RDTH_TOLT	2.59E-04	0.05	0.39	
	HS076481	31	RDTH	RDTH =f(JS + IFT)	RDTH_TOLT	6.85E-04	0.03	0.43	
	HS071704	31	RDTH	RDTH =f(JS + IFT)	RDTH_TOLT	5.98E-04	0.03	0.41	
	HS058233	31	RDTH	RDTH =f(JS + IFT)	RDTH_SELT	5.83E-05	0.05	0.08	
	HS109619	31.1	RDTH	RDTH =f(JS + IFT)	RDTH_SELT	4.88E-05	0.06	0.10	
	HS082603	31.6	RDTH	RDTH =f(JS + IFT)	RDTH_SELT	4.20E-04	0.04	0.08	
HS061187	31.8	RDTH	RDTH =f(JS + IFT)	RDTH_TOLT	1.93E-06	0.05	0.74	RDTH LO10,M0_SE10, H2O_CHA10	
HS061187	31.8	RDTH	RDTH =f(JS + IFT)	RDTH_SELT	1.83E-05	0.05	0.47		
LG15	HS132472	9.7	RDT	RDT=f(JS)	RDT_SELH	7.55E-04	0.04	0.25	
	HS070448	9.7	RDT	RDT=f(JS)	RDT_SELH	8.43E-04	0.04	0.34	
	HS066399	9.7	RDT	RDT=f(JS)	RDT_SELH	5.23E-04	0.04	0.32	
	HS122361	17.1	RDT	RDT=f(JS)	RDT_moy	1.33E-04	0.002	0.10	RDT VE10,H2O,M3,Senes
	HS058910	17.1	RDT	RDT=f(JS)	RDT_moy	2.81E-04	0.002	0.15	
	HS085663	17.1	RDT	RDT=f(JS)	RDT_moy	6.83E-05	0.002	0.17	H2O, M3
	HS122361	17.1	RDT	RDT=f(JS)	RDT_SELH	2.54E-04	0.04	0.10	RDT VE10,H2O,M3
	HS058910	17.1	RDT	RDT=f(JS)	RDT_SELH	8.98E-04	0.04	0.31	
	HS085663	17.1	RDT	RDT=f(JS)	RDT_SELH	2.28E-04	0.04	0.36	
LG16	HS060140	16.3	RDTH	RDTH =f(JS + IFT)	RDTH_SELT	7.30E-04	0.04	0.17	
	HS096335	45.5	RDTH	RDTH =f(JS + IFT)	RDTH_moy	5.69E-04	0.01	0.16	RDTH VE10 (K)
	HS149620	45.5	RDTH	RDTH =f(JS + IFT)	RDTH_moy	1.23E-04	0.01	0.26	HUILE RDTH VE09_NI CHA10 (K)
	HS120205	95.8	RDTH	RDTH =f(JS + IFT)	RDTH_TOLT	7.55E-04	0.03	0.34	

**Table IV.2 : Résumé des associations détectées avec le modèle Kais + Testeur sur les variables synthétiques.** Pour chacun des marqueurs sont indiqués : la position sur la carte consensus, leur p-value, la variance du phénotype expliqué (R2), l'effet allélique et la présence d'associations détectées variable par variable (combinaison caractère-environnement). trois marqueurs ont été positionnés par LD : HS142046, HS152545, HS096335.



Les résultats issus de l'indice de sélection combinant rendement potentiel à stress moyen et index de tolérance au stress hydrique sont quant à eux assez proches du rendement potentiel. En effet, la corrélation entre le rendement potentiel et l'index de tolérance au stress hydrique est assez faible ( $R^2 = -0.27$  pour le rendement grains et  $-0.23$  pour le rendement en huile) (Figure II.19) ce qui signifie que la PC2 traduit davantage la différence entre les génotypes liées au rendement que celle liée à la tolérance au stress JS.

La table IV.2 rassemble les statistiques pour les marqueurs dont les  $\log_{10}$  des p-values sont supérieurs à 3. Les variances phénotypiques expliquées par les marqueurs s'étendent de 0.2% à 6%, avec en moyenne 3%, ce qui est plus faible que lors des détectations univariées (variables par variables, cf section IV.3.1.5). Les marqueurs expliquant le plus de variance phénotypique et ayant les p-values les plus faibles (jusqu'à  $1.93 \cdot 10^{-6}$ ) sont ceux localisés sur le chromosome 14 et liés aux variables de stress thermique.

Les effets alléliques sont également plus faibles que pour les détectations univariées. Pour le rendement en grains potentiel ils sont inférieurs à un quintal (en moyenne 0.21). Cela pourrait signifier qu'il n'y a que peu de marge d'amélioration du rendement en conditions hydriques moyennes mais le manque de puissance de ces détectations en est aussi une cause. En effet, sur les 11 chromosomes pour lesquels des signaux intéressants ont été relevés pour le rendement potentiel (grains et huile), 9 présentent des marqueurs également associés avec d'autres caractères dont le rendement lors des tests d'associations univariées (Table IV.2). Or ces associations détectées sur quelques environnements étaient pour la plupart significatives avec un FDR de 10% et présentaient des effets alléliques plus marqués pour le rendement. Malgré la faiblesse de leur signal, les marqueurs sélectionnés ici, ont donc probablement un rôle important dans le déterminisme de ces variables synthétiques, mais il est possible qu'à chaque étape de construction de ces index, un certain « bruit » ait été introduit.

La plupart de marqueurs associés aux index de tolérance au stress hydrique (lié au RDT ou RDTH) ne sont pas associés avec d'autres variables en univarié. Par contre, un marqueur associé avec l'index de tolérance au stress thermique sur le LG14 présente également des associations avec la précocité à CHA10, le stade M0 à SE10 et le rendement en huile à LO10.



## IV.4 Synthèse des régions d'intérêt

### IV.4.1 Introduction

Les analyses univariées ont permis de mettre en évidence de nombreuses associations malgré la présence d'une forte structuration au sein du panel. Même si la plupart des zones génomiques révélées permettent d'expliquer la variation de caractères tels que la précocité, la phénologie ou l'indice foliaire, 13 chromosomes portent également des locus trouvés en association pour la productivité. Les variables synthétiques ont confirmé l'intérêt de zones « productivité » déjà détectées par les modèles univariées mais elles ont aussi révélées le rôle joué par d'autres marqueurs sur 5 chromosomes pour la tolérance au stress hydrique.

Pour les marqueurs présentant les p-values les plus significatives, l'homologue *Arabidopsis* a été identifié par BLASTX des contigs de *Helianthus annuus* portant le SNP contre la base TAIR. Les gènes identifiés et leurs annotations sont résumés dans la table IV.3. Etant donnée la faiblesse de leurs effets, les marqueurs associés risquent d'être plus ou moins éloignés du locus causal, ce qui rend cette approche imprécise. Cependant elle permet de mettre en évidence des gènes candidats potentiels pour lesquels une densification en marqueurs pourrait être envisagée. Le paragraphe suivant résumé quelques régions qui présentent un intérêt particulier pour l'amélioration de la productivité et/ou de la tolérance à la sécheresse.

### IV.4.2 Résultats

- Le LG01 comprend quatre zones indépendantes : trois zones liées avec des caractères d'indice foliaire et de phénologie et une zone liée à la productivité (RTDH) dont l'intérêt est souligné par la répétition du signal dans deux environnements différents.
- Sur le LG02, sept marqueurs associés à des caractères différents sont également indépendants. Un de ces marqueurs, associé au PMG, est situé sur un gène candidat défini dans le cadre du projet SUNYFUEL (AT1G80080) et qui code pour une protéine de type « transmembrane leucine-repeat » (ATRLP17). Cette protéine joue un

SNP	Caractère	Environnements	Chromosome	Position	Homologue	Gène candidat	Annotation
HS083828	RDTH	AI08_I	LG01	40	-	-	
HS092129	RDTH	VE10	LG01	40	-	-	
HS147170	PMG	CO08_NI	LG02	60.4	AT1G80080	oui	ATRLP17, RECEPTOR LIKE PROTEIN 17
HS061677	pmg	CO08_NI	LG03	22.6	AT2G24270	-	ALDH11A3, ALDEHYDE DEHYDROGENASE 11A3
HS105597	rdth	VE09_NI	LG03	17.5	AT2G32720	-	member of Cytochromes b5
HS089004	RDT_TOLH	-	LG03	26.8	AT2G30660	-	ATP-dependent caseinolytic (Clp) protease/crotonase family protein
HS147125	RDTH	SE10	LG05	34.3	AT1G52920	oui	G PROTEIN COUPLED RECEPTOR (ABA)
HS147124	RDTH	SE10	LG05	34.3	AT1G52920	oui	G PROTEIN COUPLED RECEPTOR (ABA)
HS147113	H2O	CO09_I	LG05	34.3	AT1G52920	oui	G PROTEIN COUPLED RECEPTOR (ABA)
HS117040	RDT	VE10	LG09	30	AT1G67430	-	Ribosomal protein L22p/L17
HS107618	RDT	VE10	LG09	30	AT1G67430	-	Ribosomal protein L22p/L18
HS090401	RDT	VE10	LG09	30.3	AT1G56110	-	NOP56-like protein
HS106242	RDT	VE10	LG09	30.3	AT5G20090	-	-
HS164549	H2O	CO09_I	LG09	32	AT5G43060	-	RESPONSIVE TO DEHYDRATION 21B, RD21B
HS164530	H2O	CO09_I	LG09	32	AT5G43060	oui	RESPONSIVE TO DEHYDRATION 21B, RD21B
HS132684	RDT	AI08_NI	LG10	51.7	AT2G47710	-	Adenine nucleotide alpha hydrolases-like superfamily protein
HS160400	RDT	AI08_NI	LG10	51.2	-	oui	
HS052625	RDTH	AI09_I	LG10	45.8	AT1G49820	-	5-methylthioribose kinase
HS146707	RDTH	AI09_I	LG10	45.6	AT3G19270	oui	CYP707A4 protein with ABA 8'-hydroxylase activity,
HS159983	RDTH	AI09_I	LG10	45.8	-	oui	
HS164332	RDTH_TOLH	-	LG10	49	AT5G19530	-	ACAULIS 5 spermine synthase
HS084642	RDT	VE10	LG12	56.8	AT5G16990	-	oxidative stress tolerance
HS077775	MOM3	VE10	LG12	69.2	AT1G24050	-	RNA-processing, Lsm domain
HS099047	MOM3	CO08_I	LG13	58.9	-	-	
HS154746	Huile	VE09_I	LG13	40	AT3G05030	oui	ATNHX2 vacuolar K+/H+ exchanger (stomatal closure regulation)
HS104806	RDTH	VE09_NI	LG13	53.3	AT3G06860	-	ATMFP2 peroxisomal fatty acid beta oxidation
HS056657	RDTH	VE09_NI	LG13	53.3	-	-	
HS057291	Huile	VE09_I	LG14	31.1	AT2G39550	-	ATGGT-1B, ABA AND AUXINE SIGNALING
HS152947	Huile	VE09_NI	LG14	27	AT2G39830	oui	DA1-RELATED PROTEIN 2 (phloem and root system development)
HS070163	PMG	CO08_NI	LG14	33.1	AT2G23420	-	nicotinate phosphoribosyltransferase 2 (NAPRT2)
HS114137	PMG	CO08_NI	LG14	34.6	-	-	
HS123988	PMG	CO08_NI	LG14	34.7	AT4G33985	-	-
HS088134	RDT	CHA10	LG17	56.3	AT1G72180	-	Leucine-rich receptor-like protein kinase family protein
HS137336	RDT	CHA10	LG17	60	AT1G12410	-	CLP PROTEASE PROTEOLYTIC SUBUNIT 2
HS111558	RDT	CHA10	LG17	56.3	AT2G45360	-	-
HS060553	RDT	CHA10	LG17	60	AT1G12410	-	CLP PROTEASE PROTEOLYTIC SUBUNIT 2
HS062427	RDT	CHA10	LG17	60.1	-	-	
HS160167	RDT	CHA10	LG17	56.3	AT5G13200	oui	GRAM domain family protein
HS147559	RDTH, RDT	CO09_I	LG17	53.3	AT1G10580	-	Transducin/WD40 repeat-like superfamily protein
HS147564	RDTH, RDT	CO09_I	LG17	53.3	AT1G10580	-	Transducin/WD40 repeat-like superfamily protein
HS147573	RDTH, RDT	CO09_I	LG17	53.3	AT1G10580	-	Transducin/WD40 repeat-like superfamily protein
HS150073	RDTH, RDT	CO09_I	LG17	53.3	AT1G10580	-	Transducin/WD40 repeat-like superfamily protein
HS150074	RDTH, RDT	CO09_I	LG17	53.3	AT1G10580	-	Transducin/WD40 repeat-like superfamily protein
HS150079	RDTH, RDT	CO09_I	LG17	53.3	AT1G10580	-	Transducin/WD40 repeat-like superfamily protein

**Table IV.3: Détails des associations discutées avec leur position sur la carte génétique, l'homologie Arabidopsis identifié, l'origine du marqueur (proviennent ou non de l'approche gène candidat) et l'annotation du gène.**

rôle dans la morphogénèse des stomates et dans le métabolisme de l'acide abscissique (Dong *et al.*, 2010).

- Le LG03 présente un nombre important de signaux lié à la productivité et à la tolérance au stress hydrique. Deux marqueurs en déséquilibre de liaison ( $r_{VS}^2 = 0.15$ ) sont associés respectivement avec PMG et RDT\_TOLH. (Un des marqueurs est associé au PMG dans un second lieu mais avec un FDR non significatif malgré sa p-value de  $4.35 \cdot 10^{-05}$ ). De plus, un troisième marqueur dont les fréquences alléliques sont moins corrélées avec celles des deux précédents marqueurs est associé avec le LAD. La sélection d'un haplotype permettrait de combiner les effets favorables pour ces trois caractères. Concernant les autres signaux détectés pour le rendement sur ce chromosome, deux marqueurs présentent des gènes aux fonctions liées même si ils ne sont pas du tout en DL. Le premier, (non significatif pour RDT avec le modèle Kais + Testeur malgré une p-value de  $8.75 \cdot 10^{-05}$ , et associé à RDT et Nbgrains avec le modèle Kais) a permis d'identifier un gène (AT5G47120) codant pour BI-1, un homologue Arabidopsis de Bax inhibitor 1, atténuant la mort cellulaire en réponse à un stress biotique ou abiotique. Le deuxième marqueur (associé à RDT avec les deux modèles de structuration) correspond à un gène (AT2G32720) codant pour un cytochrome B5. Or, il a été montré que ces deux gènes interagissaient (Nagano *et al.*, 2008) pour causer la suppression de la mort cellulaire. Un troisième marqueur associé à RDTH\_VE09\_NI correspond à un gène (AT5G54500) codant pour FQR1 (flavodoxine-like quinone reductase 1) parmi les premiers gènes dans la voie de réponse à l'auxine.

- Pour le LG05, 3 marqueurs situés sur le même gène candidat sont associés à RDTH pour deux d'entre eux et à H2O pour le troisième. Le gène sous-jacent (AT1G52920) correspond à un récepteur de l'acide abscissique. Ces conditions sont donc propices à l'utilisation d'un haplotype combinant les effets favorables.

- Comme décrit dans la section IV.3.1.5, le LG09 est caractérisé par la présence de nombreuses associations (103 associations pour 10 marqueurs) ayant des p-values fortes témoignant de la probable proximité du ou des locus causaux. A part un marqueur associé à une variable de précocité à l'extrémité du chromosome, tous les autres marqueurs positionnés entre 30 et 32 cM sont en déséquilibre de liaison et permettent d'expliquer de nombreux caractères : phénologie, précocité, foliaire et rendement sur VE10. Le rendement est encore une fois associé à la tardiveté.



Le locus causal sous-jacent est donc probablement à effet pléiotropique. Sur la zone 30-32 cM, deux marqueurs, associés à la précocité, appartiennent à un gène candidat choisi dans le cadre du projet SUNYFUEL (AT5G43060). Ce gène (RD21B) est une protéase impliquée dans la réponse à la déshydratation et au stress salin. Shindo et al (2012) ont montré que cette protéine procurait une immunité face à des pathogènes necrotrophes chez *Arabidopsis*.

- Le LG10 révèle 2 zones indépendantes pour le rendement : l'une associée à RDTH dans un environnement à potentiel (AI09\_I) et l'autre associée à RDT sur un environnement plus stressant : AI08\_NI ainsi qu'à RDTH\_TOLH (dans le modèle  $RDTH=f(JS)$ ). Le LG10 fait parti des chromosomes où nous avons observé des blocs de DL très importants notamment autour de locus qui ont été sélectionnés au cours de l'amélioration du tournesol (cf. Chapitre III). Lorsqu'on regarde les résultats du modèle Kais, davantage d'environnements ressortent associés pour le rendement avec ces marqueurs. Ces environnements qui ont disparu en ajoutant la covariable « testeur » ont pu présenter davantage de différenciation entre la performance des deux testeurs. Parmi les homologues *Arabidopsis* identifiés pour ces SNP, trois gènes candidats choisis dans le cadre des projets SUNYFUEL ou OLEOSOL apparaissent jouer un rôle important. L'un d'entre eux (AT3G19270) lié à un marqueur associé au rendement code pour une protéine cytochrome P450, bien décrite pour son rôle dans le catabolisme de l'ABA.

- Sur le LG12, six zones génomiques indépendantes ont été identifiées pour leur association avec plusieurs caractères phénotypiques dont le rendement, la durée M0M3 (liés à la tardiveté) ainsi que TOLH. Un gène impliqué dans la réponse au stress oxydatif est sous-jacent.

- Sur le LG13, chromosome caractérisé par de longs blocs de LD, la zone associée au RDTH sur un environnement est indépendante des autres marqueurs associés aux caractères de teneur en huile, de précocité, d'indice foliaire et de stade M0M3. Jusqu'à présent, les chromosomes précédents n'ont pas fait état de marqueurs associés à la teneur en huile. Sur le LG13, trois marqueurs en DL sont associés à la teneur en huile et pour deux d'entre eux également à IF20. Les génotypes aptes à maintenir leur surface foliaire active plus longtemps sont ceux dont la teneur en huile de la graine est la plus forte.





Le troisième marqueur est situé sur un gène candidat (AT3G06860) TNHX1 qui code pour un échangeur K<sup>+</sup>/H<sup>+</sup> présent dans la vacuole et impliqué dans la régulation de la fermeture stomatique. Pour le rendement, seul un SNP a permis d'identifier un homologue Arabidopsis ayant un rôle dans le métabolisme des acides gras.

- Le LG14 a permis de détecter 31 associations marqueurs-phénotype, qui d'après l'étude du DL entre ces marqueurs, peuvent être séparées en six zones. Une première zone est associée avec le PMG et la précocité : les allèles favorables au PMG sont encore une fois associés à une teneur en eau plus élevée, donc un cycle plus tardif. Une deuxième zone regroupe cinq marqueurs associés à la tolérance au stress thermique positionnés sur trois gènes candidats aux propriétés extrêmement intéressantes. En effet, deux de ces gènes candidats font partie de la même famille de facteurs de transcription « NF-YA » (NF-YA7 pour AT1G30500 et NF-YA9 pour AT3G20910). Ces gènes sont en effet impliqués dans la régulation de la réponse au stress hydrique dans la voie dépendante de l'ABA. La surexpression des gènes de cette sous unité YA génère des retards de croissance chez Arabidopsis. Quant au 3<sup>ème</sup> gène candidat (AT5G60450), il code pour un membre de la famille des facteurs de transcription ARF jouant un rôle dans le métabolisme de l'auxine. Ces marqueurs associés à la tolérance au stress thermique sont en déséquilibre de liaison avec quatre autres marqueurs associés à d'autres variables telles que la précocité, la teneur en huile et aussi le rendement en huile sur un lieu ayant un fort stress hydrique et thermique : LO10. Enfin, une zone intéressante associée à la teneur en huile permet de mettre en évidence deux gènes candidats (AT2G39550 et AT2G39830) respectivement impliqués dans la signalisation de l'ABA (Johnson *et al.*, 2005) et le développement racinaire.

- Enfin, concernant le LG17, plusieurs SNP ont été identifiés, à partir du modèle Kais, pour leur rôle dans le maintien du rendement grâce à l'évitement de la sécheresse (section IV.3.1.5). Ces 6 SNP sont positionnés dans deux gènes candidat correspondant au même homologue Arabidopsis : AT1G10580, appartenant à la famille des protéines WD40 dont le domaine est impliqué dans les interactions protéines – protéines.



## **Chapitre V Détection de QTL dans une population biparentale et comparaison avec l'approche « génétique d'association »**

### **V.1 Introduction**

Chez le tournesol, la génétique d'association n'en est encore qu'à ses débuts mais le développement des technologies à haut débit de séquençage et génotypage promettent une utilisation plus systématique.

Historiquement, l'approche la plus classique pour identifier la présence de locus impliqués dans le déterminisme des caractères quantitatifs est la détection de QTL à partir de populations biparentales. Cette approche, appelée « linkage mapping », ou cartographie de liaison, par opposition à la génétique d'association ou « Linkage disequilibrium mapping » a permis d'identifier de nombreux QTL chez la plupart des espèces d'intérêt agronomique. Malgré l'objectif commun de ces deux approches, peu d'études ont porté sur leur comparaison. Alors que la génétique d'association utilise de larges collections de génotypes souvent déjà disponibles, exploitant ainsi de nombreuses générations de recombinaison et une diversité plus large, la cartographie de liaison nécessite la création de populations dédiées et la diversité allélique est réduite à celle qui ségrége entre deux parents. La précision avec laquelle peut être identifié le QTL est un des autres arguments en faveur de l'utilisation de la génétique d'association, car l'étendue du DL y est souvent plus courte que l'intervalle de confiance d'un QTL détecté dans une population biparentale. Cependant, la présence de structuration, comme l'existence d'un groupe de mâles et d'un groupe de femelle chez le tournesol (B/R), est une des limitations principales de la génétique d'association, car elle complexifie le signal par la présence de faux positifs. Les méthodes statistiques développées pour diminuer le taux de faux positif réduisent la puissance des tests jusqu'à en masquer les locus corrélés à la structure mais réellement impliqués dans la variabilité du phénotype (faux négatif).

L'utilisation de la génétique d'association et de la cartographie de liaison en parallèle puis la mise en commun de ces résultats prend donc toute son importance. Le développement de populations biparentales dont chacun des parents appartient à une sous-population mise en évidence par l'analyse de la structure permet donc de se dégager de cet effet de structure. Grâce à la vérification des colocalisations des pics d'associations et des QTL détectés dans une population biparentale, Brachi et al (2010) ont ainsi estimé à 40% et 24% respectivement, les taux de faux positifs et faux négatifs dans le cadre d'une étude portant sur la floraison,



caractère très fortement corrélé à la structure. De plus, si la génétique d'association ne permet pas encore de pouvoir identifier des allèles rares, dont la fréquence est d'environ  $1/n$  (avec  $n$ , la taille de l'échantillon), (Morell, 2012), la cartographie de liaison est plus adaptée pour ce type d'analyse car le schéma de croisement peut artificiellement augmenter la fréquence de tel ou tel allèle.

La complémentarité de ces deux approches apparaît donc très intéressante, ne serait-ce que par la validation des zones détectées en utilisant deux dispositifs différents et aussi l'amélioration de la précision du locus causal en passant de l'échelle d'un QTL au gène candidat (Nemri *et al.*, 2010, Famoso *et al.*, 2011).

Au cours de ce projet de thèse, la plupart des caractères phénotypiques mesurés sur le panel d'association a été également enregistrée sur une des populations de référence pour le tournesol développée à l'INRA (Vear *et al.*, 2008): la population INEDI. Cette population de 273 RIL (Recombinant Inbred Lines) provient du croisement entre XRQ, lignée mainteneuses de stérilité et résistante au mildiou et PSC8, lignée restauratrice améliorée pour la résistance au Sclerotinia. Ces deux lignées parentales ont été également incluses dans le panel d'association, ce qui présente un intérêt pour confirmer la présence de QTL détectés par l'approche de génétique d'association dans la population biparentale, dans le cas où les marqueurs associés aux QTL ségrégent bien dans la population.

Comparé au dispositif mis en place pour le panel d'association où chaque lignée a été évaluée dans 17 environnements (combinaisons lieux et années), la population INEDI a été évaluée pour la plupart des caractères sur seulement 6 environnements (trois lieux à deux conditions de culture - avec ou sans irrigation - , et une seule année). Cependant, pour le caractère « date de floraison », un grand nombre de données ont été recueillies dans le cadre d'un projet Génoplante (1999-2005) avec 10 environnements supplémentaires aux six mentionnés ci-dessus. Une étude comparative équilibrée entre les deux méthodologies a pu être menée, étude faisant l'objet de l'article disponible au chapitre suivant. Contrairement à certaines publications où seules quelques régions ont été analysées via les deux approches, nous avons ici l'opportunité de faire cette étude sur l'ensemble du génome et de pouvoir ainsi avancer quelques conclusions sur l'architecture génétique du caractère.



## V.2 Etude sur le caractère floraison

### V.2.1 Résumé de l'article

L'objectif de cet article est d'identifier le polymorphisme responsable du contrôle de la date de floraison chez le tournesol en combinant une approche de génétique d'association sur le panel de 304 lignées et une approche de cartographie de liaison à partir d'une population biparentale de 273 RILs. Cet article décrit la structuration et la nature du déséquilibre de liaison au sein du panel ainsi que la méthodologie aboutissant aux tests d'associations. Nous ne développerons pas ces aspects, présentés dans la partie 2, mais seulement la méthodologie propre à la population bi-parentale et les principaux résultats de comparaison des deux approches. De la même façon que pour l'évaluation phénotypique du panel, les RILs ont été évaluées en valeur hybrides, chacune croisée avec plusieurs testeurs selon leur statut (B ou R) et le lieu. Le dispositif biparental a cependant la particularité d'inclure l'évaluation d'environ 70 RILs sur 2 testeurs par environnement pour quelques environnements ainsi qu'une évaluation *per se* (en valeurs propres). Après ajustement des données phénotypiques par un modèle spatial, la détection de QTL a été effectuée avec le logiciel MCQTL v5.2.4 (Joujon *et al.*, 2005) sur chaque combinaison caractère-environnement-testeur, appelée variable. L'étude a permis de comparer les résultats obtenus à partir de 15 variables pour le panel d'association et 23 pour la population (dont 3 variables *per se*). La détection de QTL a été menée en utilisant la méthode itérative de détection de QTL (Charcosset *et al.*, 2000) à partir d'une carte génétique publique de 517 marqueurs dont 285 SNP, version légèrement enrichie en marqueurs comparés à celle de Vincourt *et al.*, (2012). Les SNP testés en association n'étant pas tous cartographiés sur la carte publique, les marqueurs publics ainsi que les QTL ont été projetés sur une carte de 8235 marqueurs appartenant à Biogemma grâce au logiciel BioMercator v4 (Arcade *et al.*, 2004). Un test statistique décrit dans Tian *et al.*, 2011, a permis de vérifier si les colocalisations entre les marqueurs associés et les QTL de la population biparentale étaient significatives. Concernant l'étude des corrélations phénotypiques, les variables mesurées sur des hybrides impliquant le même testeur ont en général des corrélations plus fortes, qu'il s'agisse du panel ou des RILs. L'effet du testeur est donc le premier facteur qui structure la variabilité phénotypique (cf. PCA Fig.1 de l'article)





Cependant, la colocalisation entre QTL détectés sur plusieurs environnements n'est pas directement reliée au niveau de corrélations estimés par le coefficient de Pearson entre ces environnements (ou variables). Ainsi la population de RIL présente des corrélations moyennes plus faibles entre environnements que le panel et pourtant davantage de colocalisations entre environnements. Parmi les explications possibles, la structuration présente au sein du panel en est une très probable. Chaque RIL étant un patchwork de fonds génétique B ou R, la population biparentale est au contraire exempte de structuration et a donc permis de détecter davantage de QTL multilocaux.

De plus, toutes les associations détectées pour lesquelles les parents de la population sont polymorphes ont été confirmées par colocalisation, y compris celles qui n'avaient été détectées qu'avec le modèle d'association le moins complet (c'est-à-dire « Kinship » et non « Kinship + Testeur »). La cartographie de liaison présente donc un intérêt pour distinguer les vrais des faux positifs. La génétique d'association n'a semble-t-il apporté que peu de valeur ajoutée en terme de nombre de locus détectés. Cependant, elle a permis de préciser la position des locus possiblement causaux. Cinq gènes candidats impliqués dans le métabolisme de la floraison ont été ainsi mis en évidence. Enfin un QTL présent sur le LG14, identifié par l'approche biparentale ne figure pas parmi les résultats d'association. Deux hypothèses sont possibles : il peut s'agir d'un allèle que sa fréquence faible a écarté dans le processus des tests d'association, ou il peut s'agir d'un locus dont les allèles sont corrélés à la structuration B/R et qui auraient donc également été éliminé par le modèle d'association. La présence de larges blocs de DL dans cette région du LG14 lorsqu'aucune correction n'est appliquée suggère une zone génomique propre à chaque groupe de structure et donc la présence d'un faux négatif.

L'interaction génotype - environnement a été abordée dans cet article à travers l'étude des corrélations phénotypiques. Cependant, elle a été également testée avec un modèle multilocal pour le panel (les données RILs Génoplante ne le permettaient pas), ce qui a abouti à des conclusions similaires. Cette étude révèle donc que pour le caractère floraison, même si le panel d'association apporte une résolution supplémentaire, il n'a pas apporté plus de puissance de détection des polymorphismes impliqués que la population biparentale, sans doute en raison de la forte structuration présente. L'utilisation d'une population biparentale est donc très intéressante dans ce cas présent, car elle a permis de confirmer la plupart des zones et d'en détecter de nouvelles.



V.2.2 Article

## Combined linkage and association mapping of flowering time in Sunflower (*Helianthus annuus* L.)

Elena Cadic · Marie Coque · Felicity Vear · Bruno Grezes-Beset · Jérôme Pauquet · Joël Piquemal · Yannick Lippi · Philippe Blanchard · Michel Romestant · Nicolas Pouilly · David Rengel · Jérôme Gouzy · Nicolas Langlade · Brigitte Mangin · Patrick Vincourt

Received: 21 September 2012 / Accepted: 20 January 2013 / Published online: 23 February 2013  
© Springer-Verlag Berlin Heidelberg 2013

**Abstract** Association mapping and linkage mapping were used to identify quantitative trait loci (QTL) and/or causative mutations involved in the control of flowering time in cultivated sunflower *Helianthus annuus*. A panel of 384 inbred lines was phenotyped through testcrosses with two tester inbred lines across 15 location × year combinations. A recombinant inbred line (RIL) population comprising 273 lines was phenotyped both per se and through testcrosses with one or two testers in 16 location × year combinations. In the association mapping approach, kinship estimation using 5,923 single nucleotide polymorphisms was found to be the best covariate to

correct for effects of panel structure. Linkage disequilibrium decay ranged from 0.08 to 0.26 cM for a threshold of 0.20, after correcting for structure effects, depending on the linkage group (LG) and the ancestry of inbred lines. A possible hitchhiking effect is hypothesized for LG10 and LG08. A total of 11 regions across 10 LGs were found to be associated with flowering time, and QTLs were mapped on 11 LGs in the RIL population. Whereas eight regions were demonstrated to be common between the two approaches, the linkage disequilibrium approach did not detect a documented QTL that was confirmed using the linkage mapping approach.

Communicated by J. Wang.

**Electronic supplementary material** The online version of this article (doi:10.1007/s00122-013-2056-2) contains supplementary material, which is available to authorized users.

E. Cadic (✉) · Y. Lippi · N. Pouilly · D. Rengel · J. Gouzy · N. Langlade · P. Vincourt (✉)  
Laboratoire des Interactions Plantes-Microorganismes (LIPM), INRA, UMR441, 31326 Castanet-Tolosan, France  
e-mail: elena.cadic@toulouse.inra.fr

P. Vincourt  
e-mail: patrick.vincourt@toulouse.inra.fr

E. Cadic · Y. Lippi · N. Pouilly · D. Rengel · J. Gouzy · N. Langlade · P. Vincourt  
Laboratoire des Interactions Plantes-Microorganismes (LIPM), CNRS, UMR2594, 31326 Castanet-Tolosan, France

E. Cadic · M. Coque · B. Grezes-Beset · J. Pauquet  
BIOGEMMA SAS, Domaine de Sandreau, Mondonville, 31700 Blagnac, France

M. Coque · J. Piquemal  
SYNGENTA SEEDS, 12 Chemin Hobit, 31790 Saint Sauveur, France

F. Vear  
INRA, UMR 1095, Domaine de Crouelle, 234, Ave du Brezet, 63000 Clermont-Ferrand, France

P. Blanchard  
EURALIS Semences, Domaine de Sandreau, Mondonville, 31700 Blagnac, France

M. Romestant  
RAGT 2N, Site de Bourran, 12000 Rodez, France

B. Mangin  
INRA, Unité de Biométrie et Intelligence Artificielle UR875, 31326 Castanet-Tolosan, France

## Introduction

Plant breeding requires an understanding of the genetic architecture of agronomic traits, such as yield, grain quality and resistance to biotic and abiotic stresses. Many of these traits are quantitative, governed by multiple interactions of loci with effects that depend on environment. Thus, connection of marker and molecular information to phenotypes constitutes a major challenge for breeders and molecular biologists. The most frequently used approach to study this problem, initiated in the 1980's (Lander and Botstein 1989), is quantitative trait locus (QTL) mapping, referred to in this study as linkage mapping. It requires experimental populations derived from a known pedigree, the simplest of which is a cross between two parental inbred lines.

Despite the proven usefulness of this technique to identify many genomic regions involved in complex traits (reviewed in Mackay 2001), the lack of an accurate method to localize and estimate the effects of a QTL represents a serious limitation for its application to marker-assisted selection (MAS) (Holland 2004). In particular, the accuracy of QTL detection depends on the amount of genetic variability present in the experimental population.

Often presented as an alternative approach, association mapping, based on a large number of genotypes representing a broader germplasm, has several advantages over linkage mapping: (a) better resolution due to the accumulation of historical recombination events, (b) a larger number of alleles surveyed and (c) the use of already existing material, such as elite germplasm, of practical value for breeding programs (Bresghegello and Sorrells 2006). These advantages should facilitate MAS, with respect to the level of linkage disequilibrium (LD), in a population. However, association mapping has two major drawbacks compared with linkage mapping. First, it does not enable the detection of rare variants, which can reduce the power of this technique, depending on the genetic architecture underlying the trait of interest. In the case of the genetics of flowering time, two types of results have been reported. In *Arabidopsis*, genes with large effects on this trait were most commonly detected (Atwell et al. 2010). In contrast, in maize, numerous QTLs with common or uncommon genes, each explaining a small part of phenotypic variation, were detected (Buckler et al. 2009). In the majority of species, most alleles are rare (Myles et al. 2009) and can remain undetected with association mapping. A second common issue that limits the use of association mapping is the occurrence of spurious associations due to structured populations. This is particularly the case for adaptive traits such as flowering time, which is strongly correlated with population structure in *Arabidopsis thaliana* and maize (Aranzana et al. 2005; Flint-Garcia et al.

2005). Statistical methodologies have been developed to take population structure into account, but they could restrict the power of association mapping if the population is highly structured (Zhao et al. 2007). Given the complementary strengths of these detection techniques, research has been made to analyze results of association mapping and linkage mapping to combine the advantages of the two methods: resolution for the former and robustness for the latter (Nordborg and Weigel 2008).

Brachi et al. (2010) validated association peaks detected for flowering time in *Arabidopsis thaliana* when QTLs detected by linkage mapping co-localized with these peaks. Using this strategy, they were able to distinguish true from false positives and identify false negatives (causative loci lost when accounting for the population structure in the model.). Similarly, combined association and linkage mapping made possible fine mapping of QTLs in rice (Famoso et al. 2011) and in wheat (Mir et al. 2012). Association mapping methods first focused on candidate gene strategies, based on prior knowledge of the pathway controlling the trait of interest in model plants. Indeed, these strategies were guaranteed to have sufficient power, especially when the LD was high (Yan et al. 2011). Recently, the development of second generation sequencing and high throughput genotyping technologies has enabled considerable progress to be made in genetic mapping of agronomic traits. With the increasing availability of markers, the candidate gene approach has evolved towards whole genome scans, i.e., GWAS (genome wide association studies), enabling many SNPs to be queried at the same time.

Sunflower (*Helianthus annuus* L.) is a species for which massive genomic resources have been recently developed (Kane et al. 2011). It is the fourth most widely grown oilseed crop in the world. It is of major economic importance as it produces healthy oil and has low input requirements (nitrogen, water and fungicides). Sunflower production increased by 32 % over the past 20 years, reaching 32 million tons in 2010 with acreage of 23 million hectares (FAO). It has also been used as a model species to study speciation and interspecific hybridization (Rieseberg and Willis 2007). Moreover, genomic sequence of sunflower should be available soon (Kane et al. 2011). High density genetic maps are now becoming available in sunflower (Bowers et al. 2012). These maps should help genetic dissection of many important agricultural traits. Numerous linkage mapping studies for most traits, including flowering time, have been published for sunflower (detailed in this study), but only one association study is available to date, focusing on *Sclerotinia* head-rot resistance (Fusari et al. 2012).

Flowering time is a major event in the plant life cycle. This trait is controlled by both genetics and environmental

stimuli. It is related to genetic adaptation to a range of abiotic stresses (e.g., drought, light and temperature) and affects susceptibility to diseases such as *Sclerotinia* head rot. It is therefore of great interest to evolutionary biologists, eco-physiologists and breeders who need to assess flowering time in studies of crop domestication (Burke et al. 2002; Wills and Burke 2007; Baack et al. 2008; Blackman et al. 2011), response to photoperiod (Leon et al. 2001), and relations with other agronomic traits (Mestries et al. 1998; Leon et al. 2000; Mokrani et al. 2002; Bert et al. 2003; Poormohammad Kiani et al. 2009). However, the genetic architecture of this trait remains poorly understood.

In this study, we combined association and linkage mapping approaches in sunflower, based on the evaluation of genotypes per se and/or in testcrosses over several location  $\times$  year combinations. First, we evaluated the structure and LD within a large panel comprising elite lines. Then, we compared statistical models to reduce the confounding effects of associations caused by panel structures. Finally, we combined and compared linkage and association results to identify the genetic basis of flowering time.

## Materials and methods

### Plant material

The linkage mapping study was conducted on a population of 273 RILs obtained through single seed descent (to at least F8) from a cross between two INRA lines: XRQ and PSC8 (Vear et al. 2008). XRQ is a maintainer line originating from a cross between the founder USDA line HA89 and the Russian open pollinated variety ‘‘Progress’’, which confers tolerance to phomopsis and resistance to downy mildew. PSC8 was obtained from a restorer gene pool improved for *Sclerotinia* head-rot resistance by recurrent selection. Both parental lines were included in the association panel.

Association mapping was carried out on a core collection of 384 inbred lines from INRA and sunflower breeding companies, chosen for its diversity from an initial set of 752 inbred lines (Coque et al. 2008), see also [https://www.heliagene.org/Web/public/core/Core\\_collections\\_list.html](https://www.heliagene.org/Web/public/core/Core_collections_list.html). It was comprised both elite lines, parents of commercial hybrids, and lines with introgressions from several wild *Helianthus* accessions, including *H. annuus*, *H. argophyllus* and *H. petiolaris*. In this core collection, 176 lines are publicly available whereas the others are proprietary lines provided by three breeding companies: Soltis (73 lines), R2N (46 lines) and Syngenta Seeds (89 lines).

Testcross progeny were obtained by crossing association panel lines and RILs with one or two of seven testers

according to their status (maintainers of cytoplasmic male sterility [B-lines] or fertility restorers [R-lines]). These testers were chosen with the purpose of obtaining single headed, and if possible male-fertile hybrids resembling modern cultivars. They also were selected for their ability to confer standard agronomic value and resistance to major diseases to their hybrids, thus avoiding artifact effects while at the same time allowing the expression of phenotypic variability.

Testers for the RIL were CMS-PGF650 for restorer genotypes and 83HR4gms for maintainer or unbranched restorers genotypes. 83HR4gms was bred by introgressing genetic male sterility into the INRA restorer line 83HR4 (Table 1).

For the association panel, R-lines were crossed with the two CMS PET1 counterparts of B-line testers (T1 or T3) while the B-lines were crossed with two R-line testers: 83HR4gms and T2, a proprietary line carrying PEF1 cytoplasmic male sterility (Crouzillat et al. 1991) which it maintains, although it is a restorer for classical PET1 cytoplasm (Table 2). The groups of B-lines and R-lines were named B-pool and R-pool, respectively.

### Field experiments

For the RILs, testcross hybrid progeny were evaluated in four locations in 2001, three in 2002 and six in 2010. For the association panel, there were 15 location  $\times$  year combinations, (designated as ‘‘environments’’) from 2008 to 2010. While 70 RILs were evaluated on two testers per environment, the other RIL and association panel hybrids all had the same tester in any one environment. Each environment–tester combination or environment was considered to be a trait for the RIL (Table 1) or association panel (Table 2), respectively, and was analyzed separately. Each experiment was formed of blocks, with 24 or 30 entries replicated in two sub-blocks. Each sub-block was randomized separately and contained two to four check hybrids.

RILs were also evaluated per se in three additional environments, two in 2001 and one in 2004. Environments CF01 and CF04 consisted of single rows of 13–15 plants per genotype without replication, whereas there were two replications of 25 plants for CF01\_I.

In all trials, flowering time was measured as the number of days after sowing when 50 % of the plants had started anthesis.

### Phenotypic data analysis

Observations made in 2001, 2002 and 2004 on tester  $\times$  RIL combinations were first subjected to 2-way ANOVA to check statistical validity (data not shown).

**Table 1** Details on environments, testers and effective used for the 23 traits evaluated on RILs

Trait	Environment combination		Genetic profile of RILs <sup>b</sup>	Tester <sup>c</sup>	Number of RILs under evaluation
	Location <sup>a</sup>	Years			
CF01_83	CF	2001	NR, Mild.R	83HR4gms	115
CF01_PG	CF	2001	Rest.	CmsPGF650	163
CF02_83	CF	2002	NR, Mild.R	83HR4gms	115
CF02_PG	CF	2002	Rest.	CmsPGF650	155
SC01_83	SC	2001	NR, Mild.R	83HR4gms	115
SC01_PG	SC	2001	Rest.	CmsPGF650	154
SC02_83	SC	2002	NR, Mild.R	83HR4gms	115
SC02_PG	SC	2002	Rest.	CmsPGF650	154
SL01_83	SL	2001	NR, Mild.R	83HR4gms	115
SL01_PG	SL	2001	Rest.	CmsPGF650	162
RN01_83	RN	2001	NR, Mild.R	83HR4gms	115
RN01_PG	RN	2001	Rest.	CmsPGF650	163
RN02_83	RN	2002	NR, Mild.R	83HR4gms	115
RN02_PG	RN	2002	Rest.	CmsPGF650	155
AI10_I	AI_I	2010	Rest.	CmsPGF650	134
AI10_NI	AI_NI	2010	Rest.	CmsPGF650	134
GA10_I	GA_I	2010	NR	PSC8RMgms	110
GA10_NI	GA_NI	2010	NR	PSC8RMgms	110
AU10_I	AU_I	2010	Rest.	CmsXRQ	110
AU10_NI	AU_NI	2010	Rest.	CmsXRQ	110
CF01 per se	CF	2001	–	–	243
CF01_I per se	CF	2001	–	–	243
CF04 per se	CF	2004	–	–	241

The RILs were evaluated per se in 2001 and 2004, and in combination with testers in 2001, 2002 and 2010

<sup>a</sup> The locations covered the range of environments where sunflowers are cultivated in France, could be irrigated (I) or not (NI), and were designated as follows: *CF* Clermont-Ferrand (center), *SC* Longre (middle west), *SL* Baziege (south west), *RN* Villampuy (north), *GA* Gaillac (south west), *AI* Aigrefeuille (middle west), Auzeville (south west). Each trait refers to a {location × year} × tester (or per se) combination

<sup>b</sup> When evaluated in combination, a subset of RILs was chosen according to the RILs genetic profile (*NR* non-branched, *Mild.R* conferring the resistance to the race710 of downy mildew, *Rest* restauration of the male fertility)

<sup>c</sup> 83HR4gms is a modification of the restorer line 83HR4 that had been converted to genetic male sterility. PSC8RMgms is a modification of PSC8 with resistance to race 710 of downy mildew and genetic male sterility. Cms XRQ and Cms PGF650 are classical female lines

Then, to make possible comparisons between all trials, data were expressed, for each genotype, as a percentage of the check mean.

In 2010, the data collected in 15 environments for the association panel and from six environment–tester combinations for RILs were analyzed with ASReml-R (Butler et al. 2007) using the following mixed model (naïve model):

$$Y_{ijk} = \mu + G_i + b_j + c_{k(j)} + e_{ijk}$$

where  $Y_{ijk}$  is the phenotypic observation for the  $i$ th genotype in the  $k$ th sub-block of the  $j$ th block,  $\mu$  is the intercept term,  $G_i$  is the genetic effect of the  $i$ th genotype considered to be random,  $b_j$  is the effect of the  $j$ th block,  $c_{k(j)}$  is the effect of sub-block  $k$  nested in block  $j$  and  $e_{ijk}$  is the residual. Block and sub-block effects were treated as fixed.

The naïve model was enhanced in two alternative ways: (a) by the addition of random effects of row and column for the “row × column” model, or (b) including a first-order autoregressive process in the residuals to take into account autocorrelation between neighbor plots for the “ar1 × ar1” model. To compare these three models, the Aikake criterion was calculated (AIC), and its significance assessed using log ratio tests between the naïve models and the two spatial models in succession. For each trait, once the best model had been selected, the best linear unbiased predictors (BLUP) of genotypes were extracted for the next step of analysis.

For each trait, broad sense heritability ( $h^2$ ) was calculated with the following formula  $h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \frac{\sigma_e^2}{r}}$  where  $\sigma_g^2$  being the genetic variance,  $\sigma_e^2$  the residual variance, and  $r$  the

**Table 2** Details on environment, testers and effective used for the 15 traits evaluated on the association panel

Trait	Environment combination		Tester for B-pool	Tester for R-pool	Number of lines under evaluation
	Location	Years			
AI08_I	AI_I	2008	83HR4gms	T1	171
AI08_NI	AI_NI	2008	83HR4gms	T1	172
CO09_I	CO_I	2009	83HR4gms	T1	262
CO09_NI	CO_NI	2009	83HR4gms	T1	262
GA09_I	GA_I	2009	83HR4gms	T1	261
GA09_NI	GA_NI	2009	83HR4gms	T1	260
LO10	LO	2010	83HR4gms	T1	270
AI09_I	AI_I	2009	T2	T3	263
AI09_NI	AI_NI	2009	T2	T3	263
VE09_I	VE_I	2009	T2	T3	257
VE09_NI	VE_NI	2009	T2	T3	257
CA10	C1	2010	T2	T3	290
CO08_I	CO_I	2008	T2	T3	230
CO08_NI	CO_NI	2008	T2	T3	229
SE10	SE	2010	T2	T3	272

The traits are designated using the same principles as in Table 1. Within a same location  $\times$  year combination, the lines of the association panel were evaluated in testcross with the testers 83HR4gms or T2, for the lines belonging to the B-pool (See ‘‘Materials and methods’’), and with the testers T1 or T3 for those belonging to the R-pool

<sup>a</sup> The locations were designated as follows: *AI* Aigrefeuille (Middle West), *CA* Castelnaudary (South West), *CO* Cornebarrieu (South West), *GA* Gaillac (South West), *VE* Verdun (South West), *LO* Loudun (Middle West), *SE* Segoufielle (South West)

average number of replicates per genotype, which was actually close to the number of expected replicates.

Genetic and residual variances were estimated using the naïve model to compare trial precisions, independent of any specific improvements with spatial models. Phenotypic Pearson correlations and principal component analysis (PCA) were performed between traits for each type of material using R (R Development Core Team 2012).

#### Association mapping

##### Genotyping

The association panel was genotyped for 12,136 single nucleotide polymorphism (SNP) markers using the Illumina BeadXpress and Infinium platforms. Polymorphism was initially identified using three different strategies: (a) transcriptome sequencing with mRNAseq technology from samples collected from whole plants, (b) genomic sequencing of targeted genes involved in hormone signaling pathways, development, stress response or transcription, (c) non-targeted genome sequencing of gene spaces. For genotyping, DNA was extracted from young leaf tissue with a Qiagen DNeasy 96 Plant Kit using a modified protocol (Horne et al. 2004). DNA concentrations were quantified with a Quant-iT PicoGreen dsDNA assay (Invitrogen, Karlsruhe, Germany). A total of 250-ng

genomic DNA per sample was used to genotype SNP on the Illumina iScan platform (IntegraGen, Evry, France) or the Illumina BeadXpress platform (in collaboration with seed companies). Cluster positions for each marker were manually adjusted using Illumina GenomeStudio software. Heterozygous data expected to have low frequencies were considered as missing data.

A total of 8,844 high-quality SNP showing polymorphism across the association panel were retained for subsequent analysis. Eighty lines of the association panel with suspect genotypic data were discarded, giving a total dataset of 304 inbred lines for analysis.

##### Analysis of panel structure

A set of 5,923 SNP markers with less than 10 % missing data and a minor allele frequency (MAF) greater than 3 % were selected. Genotypic errors among these markers were assumed to be negligible above this threshold. Panel structure was investigated by two methods. First, a model-based approach implemented with STRUCTURE v2.2 software (Pritchard et al. 2000) was used to assign individuals to subpopulations according to correlated allele frequencies and admixture parameters. The algorithm was run for a number of subpopulations varying between one and ten. Ten replications for each subpopulation number were performed, with a burn-in time of 50,000 and 100,000



iterations. Evanno's criterion (2005) was applied to select the most likely number of subgroups.  $F_{ST}$  values similar to classical Wright's  $F_{ST}$  (1951) were used to estimate divergence between groups. Membership probabilities for each genotype in each of these subgroups were used to construct the  $Q$  matrix. Second, to perform a PCA on this dataset, each marker was centered by subtracting the average marker value across all samples and normalized by dividing by the theoretical standard deviation of the marker data at Hardy–Weinberg equilibrium (Patterson et al. 2006).

The significance of principal components was evaluated with a test based on the Tracy–Widom distribution. A PC matrix was established based on genotype coordinates of the components selected. The relatedness between all pairs of individuals were estimated using the AlikeIn State estimator (AIS) in Cocoa software (Maenhout et al. 2009), which comprised the  $K_{ais}$  matrix. To assess the relative performance to correct panel structure,  $Q$ , PC and  $K_{ais}$  matrix values were successively specified as covariates in association models.

In addition, a structure effect denoted ‘‘Tester’’, due to the status (B/R) of the lines, was included in model comparisons. As tester lines used to obtain hybrids were different between B and R lines, the Tester effect did not dissociate the effects of B or R status from the particular genotypic effect(s) of the tester(s). In each model, the structure was considered to exert fixed effects, and the genetic background captured by  $K_{ais}$  matrix was considered to be random, following the recommendations of Yu et al. (2006). A summary of all the models compared is presented in Table 3. Bayesian information criterion (BIC) and the  $p$  value of the significance of the fixed effects were used to determine the best-fitting model for each trait. The ability of each model to control for type I errors was

compared by examining  $p$  values of association tests on a set of 1,000 random markers using ASReml-R. Finally, Student's test in R was used to check whether the mean of each phenotypic trait differed significantly between subpopulations.

#### Linkage disequilibrium

The resolution of association mapping studies depends on the pattern of linkage disequilibrium (LD). For LD estimation, we used a set of 1,874 SNP mapped on the proprietary BIOGEMMA consensus map and genotyped on the association panel, with a MAF of 5 % and missing data of <10 %. LD was calculated for each chromosome between all pairs of markers using classical statistics ( $r^2$ , squared correlations, between two loci, here in their haploid state) and a new measure correcting for biases caused by structure and relatedness between individuals obtained:  $r_{vs}^2$  (Mangin et al. 2011). Classical and new LD statistics were plotted as a function of genetic distance to estimate LD decay per chromosome, using Hill and Weir's model (1988), with a threshold of 0.2.

#### Association tests

Association mapping was based on a set of 6,645 SNP, including markers used for structure estimation and markers localized in candidate genes. This sample was taken from the validated set of 8,844 SNP after removing data containing more than 10 % missing observations and MAF lower than 5 %. Association between single loci and traits was carried out in Emma (Kang et al. 2008) using the two mixed models that correct for genetic relatedness between lines: ‘‘ $K_{ais}$ ’’ and ‘‘ $K_{ais} + \text{Tester}$ ’’. The full statistical model is:

**Table 3** Summary of models tested for association detection

Model	Description	Covariate specification
Naïve	No correction for population structure	–
Fixed effects		
$Q$	Population structure inferred by STRUCTURE	Proportion of genome assigned at each group (g1, g2, g3)
PC	Population structure resulting from PCA	Coordinates on principal components
Tester	Structure of breeding pools (B-pool, R-pool)	Binary (0/1)
Random effects		
$K_{ais}$	Relatedness as estimated by alikeness in state coefficient	Matrix of variance–covariance proportional to the pairwise AIS matrix
Mixed		
$K_{ais} + Q$	Mixed model with population structure as fixed effects and relatedness as random effect	
$K_{ais} + \text{PC}$		
$K_{ais} + \text{Tester}$		

$$G_i^{BLUP} = \sum_c X_{ic} a_c + M_{il} \theta_l + u_i + e_i$$

$G_i^{BLUP}$  is BLUP for  $i$ th hybrids,  $X_{ic}$  is tester category (0–1),  $a_c$  is effect of tester category  $c$ ,  $M_{il}$  is genotype of the  $i$ th hybrid at locus  $l$ ,  $\theta_l$  is effect of locus  $l$ .  $a_c$  and  $\theta_l$  are considered to be fixed effects.  $u_i$  is the random polygenic effect modeling genetic relatedness with  $\text{Var}(u) = \sigma_u^2 K_{\text{ais}}$  where  $K_{\text{ais}}$  is an AIS matrix and  $\text{Var}(e) = \sigma_e^2$ .

A false discovery rate (FDR) (Benjamini and Hochberg 1995) was applied on  $p$  values to correct for multiple testing. Variance explained by each marker was estimated using the  $R_{LR}^2$  statistics described in Sun et al. (2010).

### Linkage mapping of QTLs

A consensus genetic map was built using the method explained in Vincourt et al. (2012) and QTLs detected by linkage mapping in the RIL population were designated LM-QTL in this study. This consensus map and the corresponding map for the RIL population were produced from 619 and 517 public markers, respectively, of which respectively 345 and 285 were SNP (Supplementary File 1 also available at [https://www.heliogene.org/Web/public/Consensus\\_INEDI\\_FUxPAZ2\\_V1/mapping\\_INEDI\\_FUxPAZ2\\_2012-07.html](https://www.heliogene.org/Web/public/Consensus_INEDI_FUxPAZ2_V1/mapping_INEDI_FUxPAZ2_2012-07.html)). QTLs were detected for each trait (environment-tester combination) with MCQTL v.5.2.4 (Jourjon et al. 2005) using iterative composite interval mapping (Charcosset et al. 2000) with the model:

$$G_i^{BLUP} = \mu + \sum_{l=1}^n \sum_{g=1}^2 P(G_i^l = g | M_i) \beta_g^l + e_i$$

where  $G_i^{BLUP}$  is the BLUP for the  $i$ th RIL,  $\mu$  is the global mean,  $P(G_i^l = g | M_i)$  is the probability of individual  $i$  having genotype  $g$  at QTL or cofactor locus  $l$  given the whole marker information denoted  $M_i$ ,  $\beta_g^l$  is the mean of genotype  $g$  at locus  $l$  and  $e_i$  is the residual.

The statistical significance of QTLs was assessed using the MCQTL test, which is equal to  $-\log(p \text{ value } (F \text{ test}))$ , as described in the MCQTL version 5 reference manual (<http://carlit.toulouse.inra.fr/MCQTL/>). A genome-wide type I error rate of 0.05 was applied, estimated after performing 1,000 permutation tests for each trait. QTL confidence regions were determined using a two-LOD support interval.

### Overlap between linkage and association signals

An 8,235 marker proprietary consensus map developed by BIOGEMMA (unpublished data) served as a reference to project the 517 public markers and the detected QTLs using BioMercator v4 (Arcade et al. 2004). When possible,

associated markers were located on this map. For those unmapped, linkage disequilibrium ( $r_{vs}^2$ ) with all positioned SNP was calculated to place the markers. Unmapped SNP were assigned to the same position as the mapped SNP that was in maximum LD if the LD statistic was above a threshold of 0.1. The overlap between detected QTLs and association signals was tested statistically according to Tian et al. 2011. The total genome distance covered by a QTL over the genome size (in centiMorgans, cM) was computed to determine the probability of an SNP falling into a QTL support interval. When several QTLs overlapped, the largest interval was chosen. The hypothesis that associated SNP was placed with a larger probability in a QTL region than was expected by chance was tested using a binomial distribution.

In addition, LD statistics  $r_{vs}^2$  between markers positioned in regions where QTLs overlapped with associated markers were used to build LD heat maps from a modified code of `snp.plotter` function (Luna and Nicodemus 2007).

## Results

### Phenotypic data analysis

The period from sowing to flowering time was measured on the association panel and RILs in a variety of environments. As indicated above, flowering time in each environment (for the association panel) or each environment–tester combination (for the RIL population) was considered to be a separate trait. In a first step, statistical analyses were made on the 15 traits obtained on the association panel and the six traits obtained in RILs in 2010.

Statistics for the naïve model, i.e., including only fixed blocks and sub-block effects with a random genotype factor, were computed (Supplementary File 2). Genotypic variance differed significantly from zero for all traits. Broad sense heritability ranged from 0.55 to 0.96, with a lower mean (0.68) for the RILs than for the association panel (0.84). Spatial models displayed a significantly better fit than the naïve model with respect to the AIC criterion. The model “ar1 × ar1” was more appropriate in 13 environments, while the “row × column” model performed better in six environments and the naïve in one. However, BLUP extracted from these three models were also highly correlated (data not shown). We selected BLUP from the best model to perform further analysis.

Correlations between the 15 association panel traits were all significant ( $p < 0.001$ ; Supplementary File 3). A higher mean correlation between traits derived from the environments allocated to testcrosses made with the same pair of testers was observed: 0.66 and 0.72 for the group of 83HR4gms/T1-related traits and for the group of T2/T3-

related traits, respectively, to be compared with 0.49 between the two groups. This result was also illustrated by the PCA (Fig. 1). We still observed a large amount of variability in flowering time occurrence and distribution within each group. For example, T2/T3 group showed flowering time differences spanning 24 days.

Correlations between traits recorded on RILs (complete set of 23 traits, Supplementary File 4) were less significant. The mean correlation coefficient between traits recorded on hybrids involving the tester 83HR4gms was 0.51, compared to 0.56 for the tester CmsPG650. Similarly to what we observed for the association panel, on the second principal axis, PCA revealed a clear distinction between traits derived from different testers, with 83HR4gms- and CmsPG650-related traits being the most distant.

Despite of the overall good correlation between traits, we chose to conduct association and linkage mapping for each trait independently to capture specific interactions associated with a particular tester or environment.

### Population structure

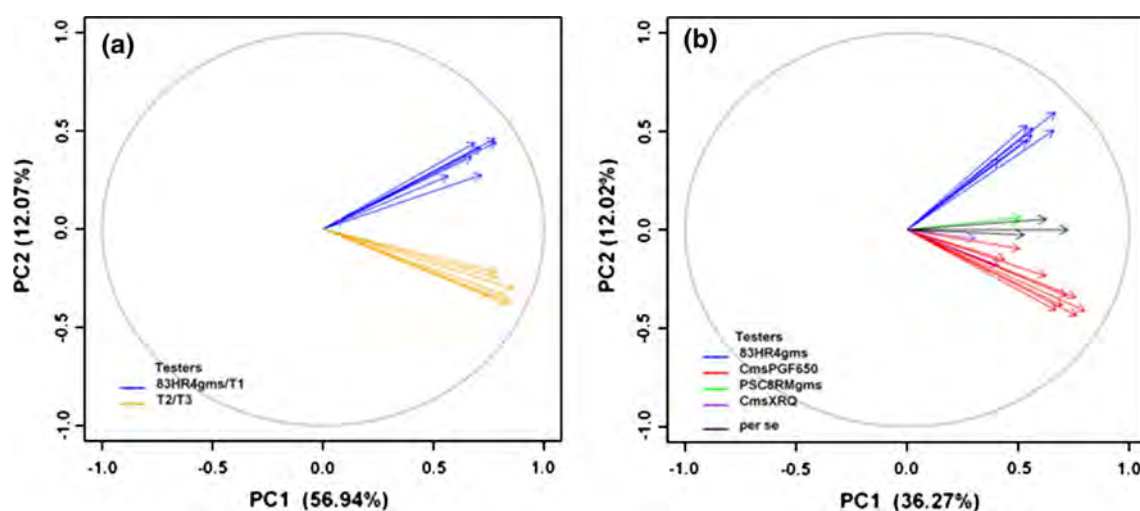
Using the Evanno's criterion (delta Evanno), the model-based STRUCTURE approach distinguished three probable groups, g1–g3, in the panel, as illustrated in Fig. 2a, b. The second criterion, i.e., the distribution of the log likelihood of the data, was not very meaningful because it did not reach a plateau. Two of the three groups exhibited by STRUCTURE were made up of wild introgression lines, belonging either to a set of 29 B-lines for the “g1 group” or to a set of 36 R-lines for the “g2 group”. A total of 27 over 29 lines were assigned to g1, with a mean percentage of 0.98, and 22 lines over 36 were assigned to g2, at a

percentage of 0.90. The third group (“g3” group) contained a majority of public B-lines. The g1 and g2 groups presented higher  $F_{ST}$  values (0.57 and 0.40, respectively) than g3 (0.07). The groups inferred by STRUCTURE are also highlighted in the PCA (Fig. 2c), where in addition to g1 and g2, g3 appears clearly in the PCA as a dense block of related individuals. For 151 inbred lines, there was no evidence of clear assignment to one of these three groups at a threshold of 0.80. The set of these 151 lines will be named g4 thereafter for convenience.

R/B line divergence was more obvious when using PCA analysis (Fig. 2c) than in the STRUCTURE analysis. Based on the Tracy–Widom statistics (Patterson et al. 2006), the first three principal components were considered to be significant and explained 13.21 % of the total variability. The first principal component, explaining 5.91 % of the variability, separated the B-pool on the right side with the g1 group on top and the R-pool on the left side with the g2 group on top.

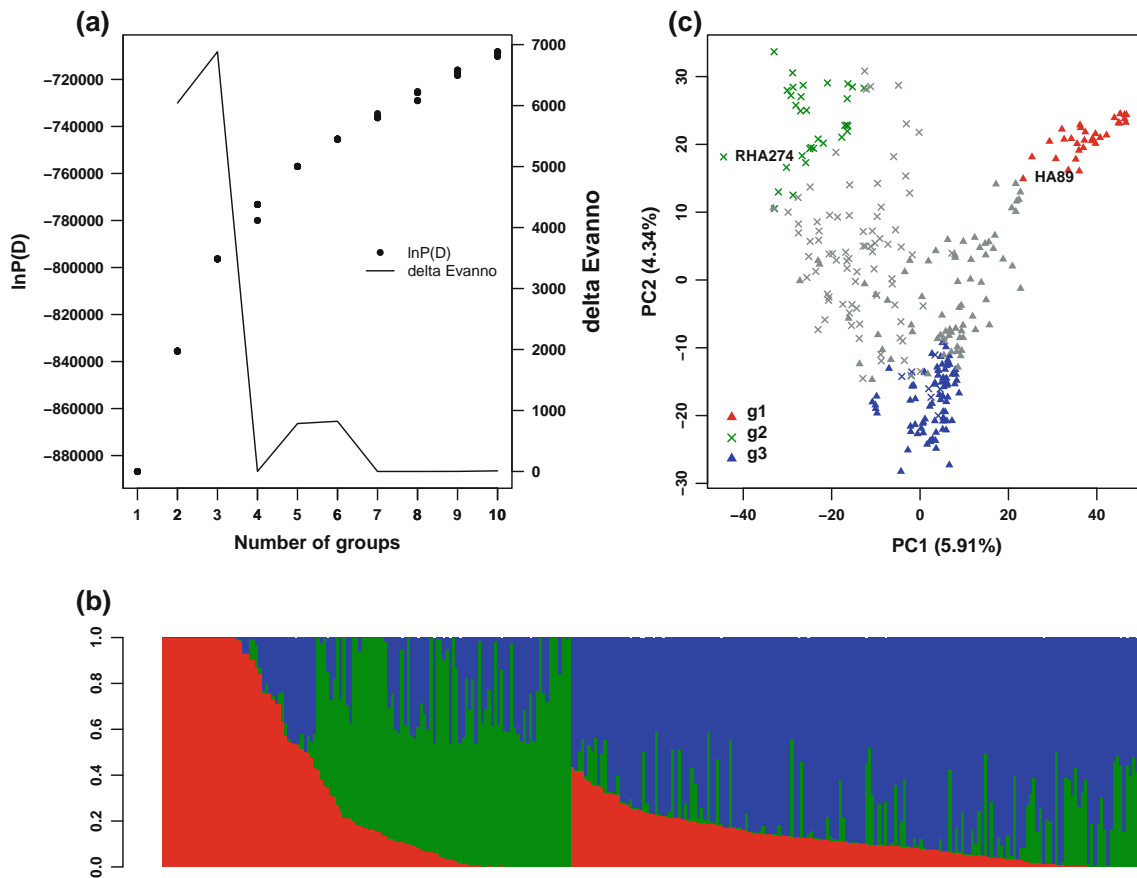
### Models comparison for association mapping

The majority of traits presented a mean significant difference between R- and B-lines, as well as between groups detected by STRUCTURE, highlighting the need for structure correction in association tests. Thus, a total of eight models were compared for their ability to correct stratification for each trait (Table 3). In the first step, marker information was not taken into account when exploring BIC criteria (Table 4) and  $p$  values of fixed effects (Supplementary File 5). Among the eight models compared, two reached the lowest BIC for most of the traits, including the “ $K_{ais}$ ” model for seven environments



**Fig. 1** Principal coordinate plots for the flowering time phenotypic traits recorded on the association panel (15 traits, **a**) and on the RIL population (23 traits, **b**). Percentages in parentheses refer to the

proportion of variance explained by first and second principal coordinates (color figure online)



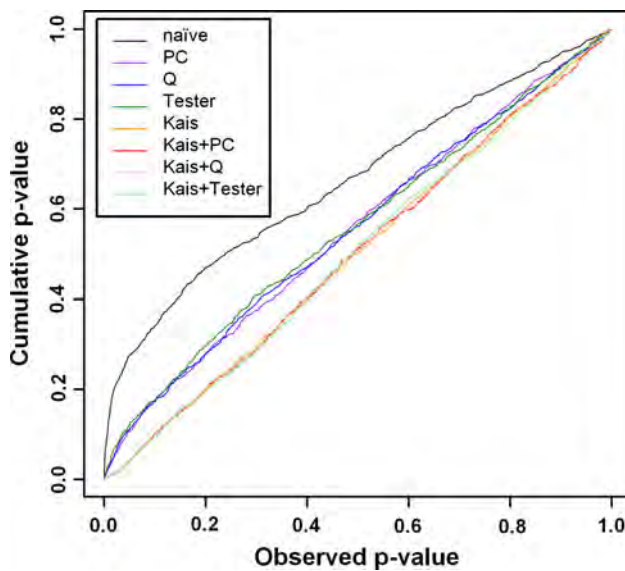
**Fig. 2** Stratification of the 304 association panel lines. **a** Log-likelihood and delta Evanno statistics for a number of putative populations ranging from 1 to 10, **b** STRUCTURE output for three groups, **c** top two principal components in PCA. Percentages in

parentheses refer to the proportion of variance explained by the principal coordinate. Symbols represent the two breeding pools (x for R lines and filled triangle for B lines). In **b** and **c**, each genotype is colored according to its STRUCTURE group (color figure online)

**Table 4** Comparison of Bayesian information criteria (BIC) among eight models

Traits	Naïve	Fixed effects			Random effects $K_{ais}$	Mixed effects		
		$Q$	PC	Tester		$K_{ais} + Q$	$K_{ais} + PC$	$K_{ais} + Tester$
AI08_I	528.35	526.73	529.14	515.93	513.61	523.18	528.17	<i>511.46</i>
AI08_NI	410.47	417.78	422.13	410.24	<i>409.08</i>	419.94	425.35	412.02
AI09_I	1,283.27	1,258.12	1,260.24	1,249.42	1,227.81	1,238.26	1,241.99	<i>1,221.57</i>
AI09_NI	1,221.61	1,194.95	1,198.11	1,184.06	1,145.96	1,156.85	1,161.30	<i>1,138.55</i>
CO08_I	1,038.25	1,031.42	1,036.03	1,032.15	<i>1,003.58</i>	1,014.32	1,019.02	1,008.59
CO08_NI	1,051.91	1,049.51	1,054.98	1,054.94	<i>1,018.95</i>	1,030.00	1,034.65	1,024.43
CO09_I	888.63	889.51	894.05	881.06	<i>877.57</i>	888.51	893.89	879.44
CO09_NI	941.38	926.07	929.03	918.22	916.22	926.24	930.62	<i>915.62</i>
GA09_I	988.50	947.65	946.65	937.98	941.09	947.65	949.88	<i>933.47</i>
GA09_NI	946.87	907.37	909.53	902.56	901.20	908.62	912.24	<i>896.73</i>
VE09_I	1,155.77	1,165.35	1,169.99	1,161.12	<i>1,120.86</i>	1,131.60	1,137.58	1,126.38
VE09_NI	1,180.22	1,189.04	1,192.37	1,185.16	<i>1,173.77</i>	1,185.10	1,190.19	1,179.23
CA10	1,409.67	1,411.75	1,416.63	1,399.70	<i>1,343.59</i>	1,353.56	1,359.90	1,344.40
LO10	1,081.28	1,038.51	1,042.78	<i>1,023.55</i>	1,042.25	1,045.34	1,049.39	1,029.12
SE10	<i>1,072.14</i>	1,080.56	1,085.58	1,074.08	1,075.05	1,086.33	1,091.69	1,079.79

Values in italics correspond to the lowest BIC for each trait



**Fig. 3** Cumulative  $p$  value distribution of the association scan over a set of 1,000 random SNPs. The same eight models as in Table 4 were compared for the trait AI09\_I (color figure online)

and the “ $K_{\text{ais}} + \text{Tester}$ ” model for six environments. BIC values were found to be quite similar between these two models.

We compared  $p$  values of fixed effects in six out of the eight models (“ $K_{\text{ais}}$ ” and naïve models were excluded). Significant structure effects were observed for most of the traits in the fixed models. When relatedness was taken into account in these models,  $p$  values of STRUCTURE and PCA covariates became non-significant (except for LO10). In contrast, the Tester effect remained significant for six traits, including those best fitted by the “ $K_{\text{ais}} + \text{Tester}$ ” or “Tester” models, according to the BIC criterion. For CO09\_NI, the Tester effect was not significant, although this trait showed best fit with the “ $K_{\text{ais}} + \text{Tester}$ ” model. These results show that principal components and STRUCTURE covariates were not needed to correct for structure effects, as kinship probably retained a large amount of this information.

In the second step of analysis, we incorporated a set of 1,000 SNP into the investigated models to search for an excess of significant associations over those predicted by the null hypothesis. The quantile–quantile (Q–Q) plots looked very similar between traits; a representative example is presented in Fig. 3. For the naïve association model, the rate of false positives was clearly inflated, suggesting that correction was necessary. Models specifying structure effects performed better, and only models including kinship matched the diagonal, showing good control of false positives. The above analysis demonstrates that both the “ $K_{\text{ais}}$ ” and “ $K_{\text{ais}} + \text{Tester}$ ” models, considered to be the best models according to BIC and  $p$  value

criteria, gave similar reductions in false positives. As a consequence, we decided to conduct association tests using these two models.

#### LD estimation

We first investigated the pairwise LD for each chromosome in the entire panel, with (Mangin et al. 2011) or without correcting for B/R line structure and kinship confounding effects. Mean LD decay, defined by the Hill and Weir model (1988) with a threshold of 0.20, was estimated across each LG. The value decreased from 0.41 cM without correction to 0.14 cM with kinship and structure correction. As illustrated in Fig. 4a for LG08, long distance pairwise LD (over 10 cM) was not maintained after correction. However, we observed considerable differences in LD decay, ranging from 0.08 to 0.26 after correction, between LGs. LG08 and LG10 presented a specific pattern, with high LD values when no correction was applied (Fig. 4b).

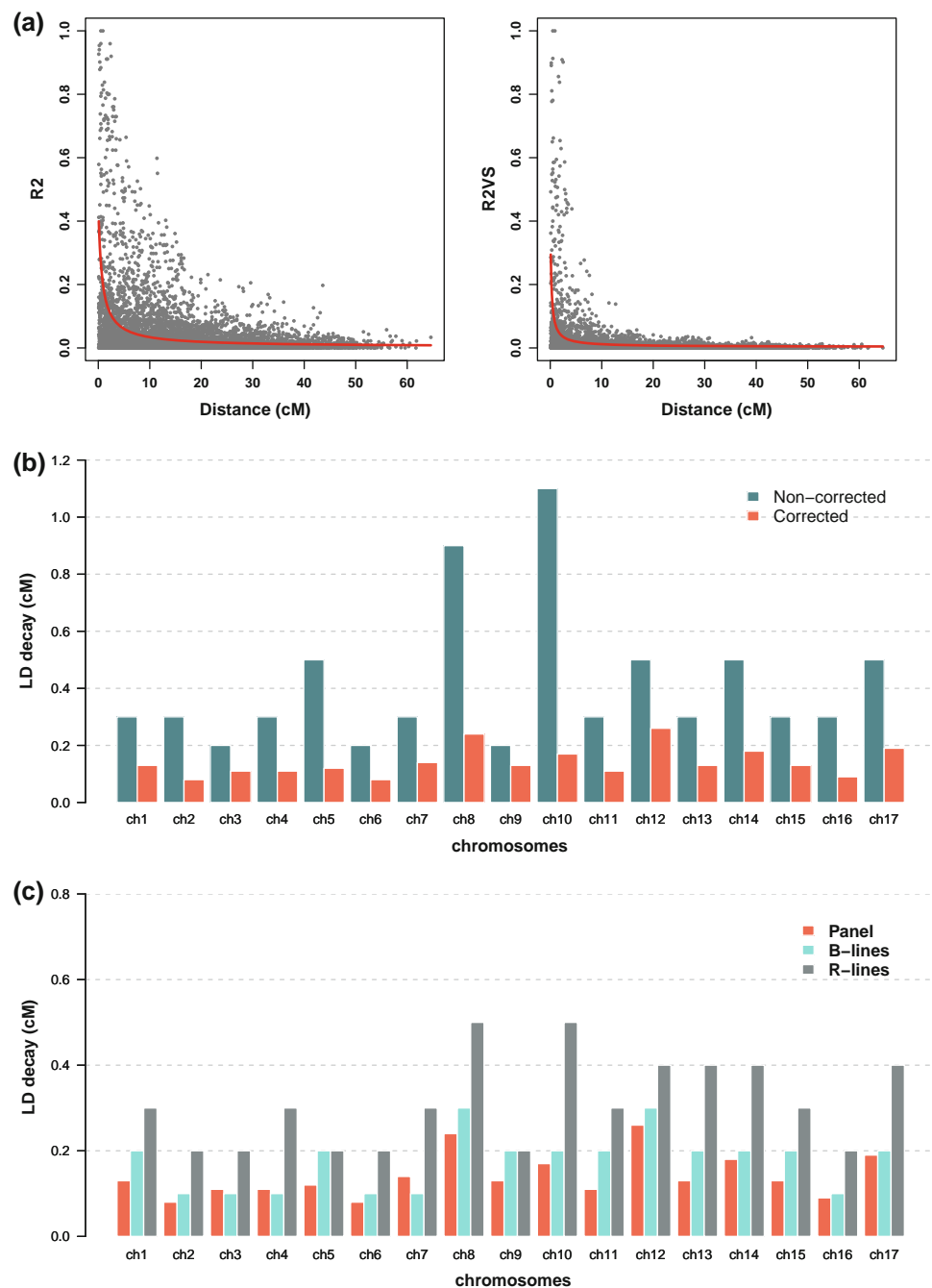
In a second step, we computed the corrected LD statistics accounting only for kinship effect:  $r_{\text{vs}}^2$ . This was done separately for the R-pool (121 accessions) and the B-pool (183 accessions), as maintainer and restorer lines are considered to belong to distinct breeding pools. For most chromosomes, the B-lines presented a mean decay of LD that was similar to that of the entire panel. In contrast, the R-lines showed higher LD values, with a mean decay of 0.31. LGs having the largest LD differences between the two breeding pools, with a higher value for R-lines pool, included chromosomes 8, 10, 13, 14 and 17 (Fig. 4c).

#### Mapping results

##### Linkage mapping

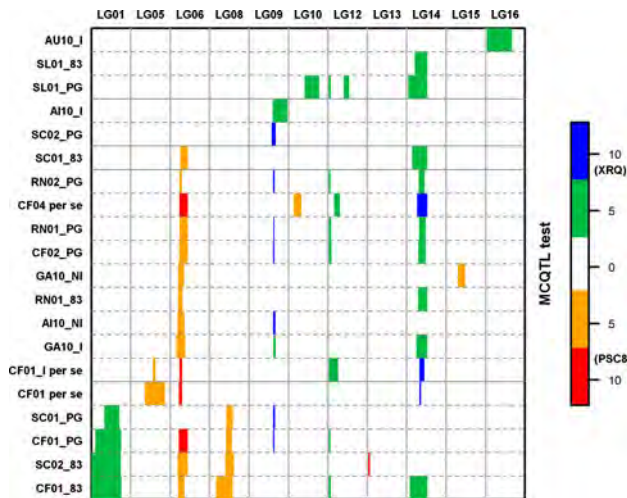
QTLs were detected for 20 out of 23 traits. The QTLs accounted for 6–29 % of the phenotypic variation, with an additive effect varying from 0.1 to 5.0 days (Detailed statistics in Supplementary File 6, linkage map with QTL available at [https://www.heliagene.org/Web/public/linkage\\_mapping\\_FT\\_INEDI.html](https://www.heliagene.org/Web/public/linkage_mapping_FT_INEDI.html)). Considering the 2001 and 2002 environments in which hybrids were tested, the crosses with CmsPGF650 led to the identification of more QTLs (23) than for those with 83HR4gms (13). For environments where the two testers were used, we identified QTLs on both of them, especially on linkage groups 6 (LG06) and LG14 (Fig. 5). In contrast, several regions were only detected with one tester. For example, the LG09 region was identified in material crossed to the tester CmsPGF650. This region, with an average confidence interval of 6.60 cM, was highly significantly associated with nine traits (average MCQTL test = 9.44 and explained variance of 24 %), with allelic effects reaching

**Fig. 4** **a** LD statistics plotted over genetic distance (cM) on LG08: classical LD (*left*),  $r_{vs}^2$  (LD corrected from relatedness and B/R line structure, *right*); **b** distribution of classical LD and  $r_{vs}^2$  statistics across chromosomes; **c** distribution of  $r_{vs}^2$  across chromosomes for the entire panel and for each breeding pool (B-lines and R-lines) (color figure online)



5 days. In addition to this QTL, three regions were particularly highlighted by overlapping of QTL support intervals for a large number of traits. LG06 contained a cluster of QTLs detected in 12 environments overlapping in a 21-cM support interval. A cluster of QTLs corresponding to 10 environments and overlapping within a 33 cM support interval was found on LG14. LG12 presented a smaller confidence region of 14 cM, where QTLs for seven environments were mapped. While the allele conferring late flowering were derived from XRQ for the QTL located on

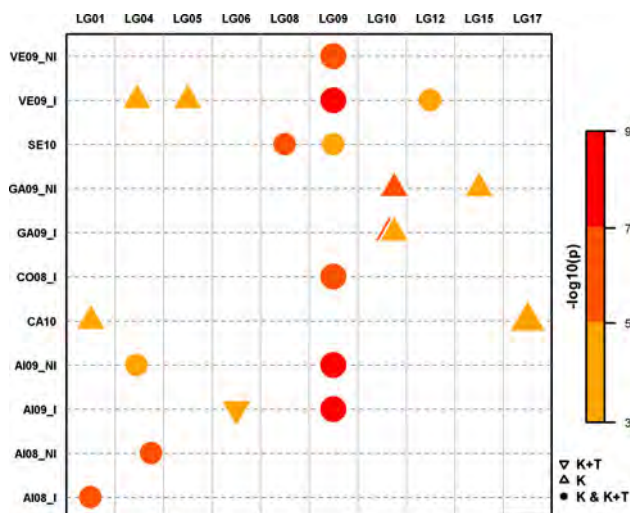
LG09, LG12 and LG14, it came from PSC8 for the QTL located on LG06. Fourteen QTLs were detected in environments where RILs per se were evaluated. Overall, they presented a higher significance level than the rest of QTLs detected (MCQTL test = 8.00 vs. 5.3) but with a lower additive effect (0.12 vs. 2.10 for the QTL detected on hybrids). Compared to results obtained for hybrids, new regions appeared to be involved in the flowering trait evaluated on RILs per se on LG05 and LG10, while regions on LG06, LG08, LG12 and LG14 were confirmed.



**Fig. 5** Heat map of detected QTLs. Only chromosomes (*columns*) and traits (*row*) where QTLs were detected at a threshold of 3.80 are represented. A *color scale* is used to indicate QTL significance (MCQTL test). Positive values (*orange and red*) denote a later occurrence of flowering time for RILs carrying PSC8 alleles (color figure online)

#### Linkage disequilibrium mapping

Figure 6 presents the results of marker trait association analysis for “ $K_{ais}$ ” and “ $K_{ais} + Tester$ ” models and their localization on the proprietary map. Eleven of the 15 traits presented significant associations at the FDR threshold of 10 %. Most were detected using both models. However, several signals found associated with a trait using the “ $K_{ais}$ ” model were not identified using the “ $K_{ais} + Tester$ ” model.



**Fig. 6** Heat map of significantly associated markers mapped (or localized with LD) on the consensus map for each trait and both tested models (“ $K_{ais}$ ” or “ $K_{ais} + Tester$ ”). A *color scale* is used to indicate SNP significance based on  $p$  values ( $-\log_{10}(p)$  value)) (color figure online)

To summarize, 11 regions on 10 LG were found to be associated with flowering time. One of these regions, on LG09, was noticeable due to the number of traits (6) for which an association was detected. In this region, very low  $p$  values were found for some of the markers. For example, SNP HS117598 presented a  $p$  value of  $7.65 \times 10^{-10}$  and additive effect of 3 days on flowering time in the AI09\_I environment. Moreover, the markers spanned an interval greater than 2 cM, thus showing the linkage disequilibrium in this region. The allelic effect was found in the same direction for each trait. Apart from this region and the three markers mapped on LG01, 04 and 10, respectively (detected in two environments each), all association signals were specific to one environment. We also observed poor consistency in mapping results between highly correlated environments, such as those differing only by irrigation treatment. Mapping results were not consistent between traits obtained with 83HR4gms/T1 and those obtained with T2/T3, as only one marker on LG01 was detected in both groups of traits. Details of MAF,  $p$  values, allelic effects and variance explained for each SNP detected (range 4.5–13.3 %) are provided in supplementary data (Supplementary File 7).

#### Overlap between association signals and LM-QTL

SNP identified using association mapping were compared with the positions of QTLs detected in the RIL population. For this purpose, a consensus map was built by projecting the LM-QTLs detected on the public map onto the proprietary map. The significance of overlapping between association peaks and LM-QTLs was assessed as follows. We first identified independent SNP among the 27 SNP found associated with traits by calculating LD  $r_{vs}^2$  (corrected statistics) for each pair of markers. Among the 13 SNP considered to be independent  $r_{vs}^2$  (threshold of 0.1) and mapped by recombination or LD on the consensus map, nine were positioned in LM-QTL support intervals. Binomial tests proved that the overlap observed was significant ( $p = 3.49 \times 10^{-6}$ ), compared to the probability of an SNP falling into an LM-QTL region by chance (0.17).

A total of eight chromosomes carried QTLs on which associated SNP were also detected (Table 5). Figure 7 describes the pattern of LD combined with associated  $p$  values for two regions of interest: LG09, which displays highly significant associations consistent across environments, and LG10, on which one marker is positioned in a candidate gene for flowering time (detailed in the discussion). On the latter chromosome, two markers were significantly associated with this trait. One of these markers was not localized in the QTL regions (Fig. 7a) but was in LD to the second marker.

**Table 5** Significant markers detected with association mapping and mapped within the QTL regions found by linkage mapping

LG	Map position	Marker	QTL interval on the consensus map	Arabidopsis homolog locus	Arabidopsis homolog gene name (if any)	Arabidopsis homolog TAIR description
LG01	33.5	HS136120	−6.00–72.71	AT5G18120	APRL7	Encodes a protein disulfide isomerase-like (PDIL) protein, a member of a multigene family within the thioredoxin (TRX) superfamily. This protein also belongs to the adenosine 5′-phosphosulfate reductase-like (APRL) group.
LG05	44.8 <sup>a</sup>	HS107108	34.22–66.11	AT3G17590	BSH	Encodes the Arabidopsis homolog of yeast SNF5 and represents a conserved subunit of plant SWI/SNF complexes.
LG06	32.2	HS113607	18.8–35.7	Not found		
LG08	29.0	HS097037	−20–30	AT1G75560		Zinc knuckle (CCHC-type) family protein
LG09	31.3	HS117040	27.08–35.97	AT1G67430		Ribosomal protein L22p/L17e family protein
LG09	31.6	HS090401		AT1G24030		Protein kinase superfamily protein
LG09	31.6	HS117598		AT1G67580		Protein kinase superfamily protein
LG09	31.8	HS095606		AT5G50260	CEP1	Cysteine proteinases superfamily protein
LG10	53.2	HS061549	41.12–66.8	AT3G63010	GID1B	Encodes a gibberellin (GA) receptor ortholog of the rice GA receptor gene (OsGID1). Has GA-binding activity, showing higher affinity to GA4. Interacts with DELLA proteins in vivo in the presence of GA4.
LG10	57.2	HS073886		AT5G47780	GAUT4	Encodes a protein with putative galacturonosyltransferase activity.
LG12	34.8	HS067214	45–66.8	AT3G14310	PME3	Encodes a pectin methylesterase, targeted by a cellulose binding protein (CBP) from the parasitic nematode <i>Heterodera schachtii</i> during parasitism.
LG15	42.2 <sup>a</sup>	HS057257	36.48–51.04	Not found		

Only markers corresponding to different genes are indicated. Map positions refer to the consensus map built by projection of the public map onto the proprietary map. The putative *Arabidopsis thaliana* homologs were obtained on TAIR by blasting (BLASTX) the *Helianthus annuus* contig sequence carrying the SNP

<sup>a</sup> These markers were mapped on the consensus map based on their LD with mapped markers

## Discussion

### Panel structure and breeding history

As a prerequisite for association mapping, panel population structure was assessed using a PCA and a Bayesian model in STRUCTURE software. Results with the two analyses were quite similar and confirmed by the significant correlation between the *Q* and *P* matrices (data not shown).

The panel appeared to be made up of three groups (Fig. 8), the two most divergent groups g1 and g2 contained respectively B and R lines. The proximity among the g1 lines in one hand and among the g2 lines in other hand is probably due to founder effect which arose from the introgression of wild *Helianthus* accessions into an elite B line and an elite R line, respectively. This suggests that it should be possible to enhance genetic variability in both B and R gene pools by increasing the use of wild *Helianthus* accessions in breeding programs. It should be pointed out that the two well-known USDA lines HA89 and RHA274 were located in the g1 and g2 groups, respectively.

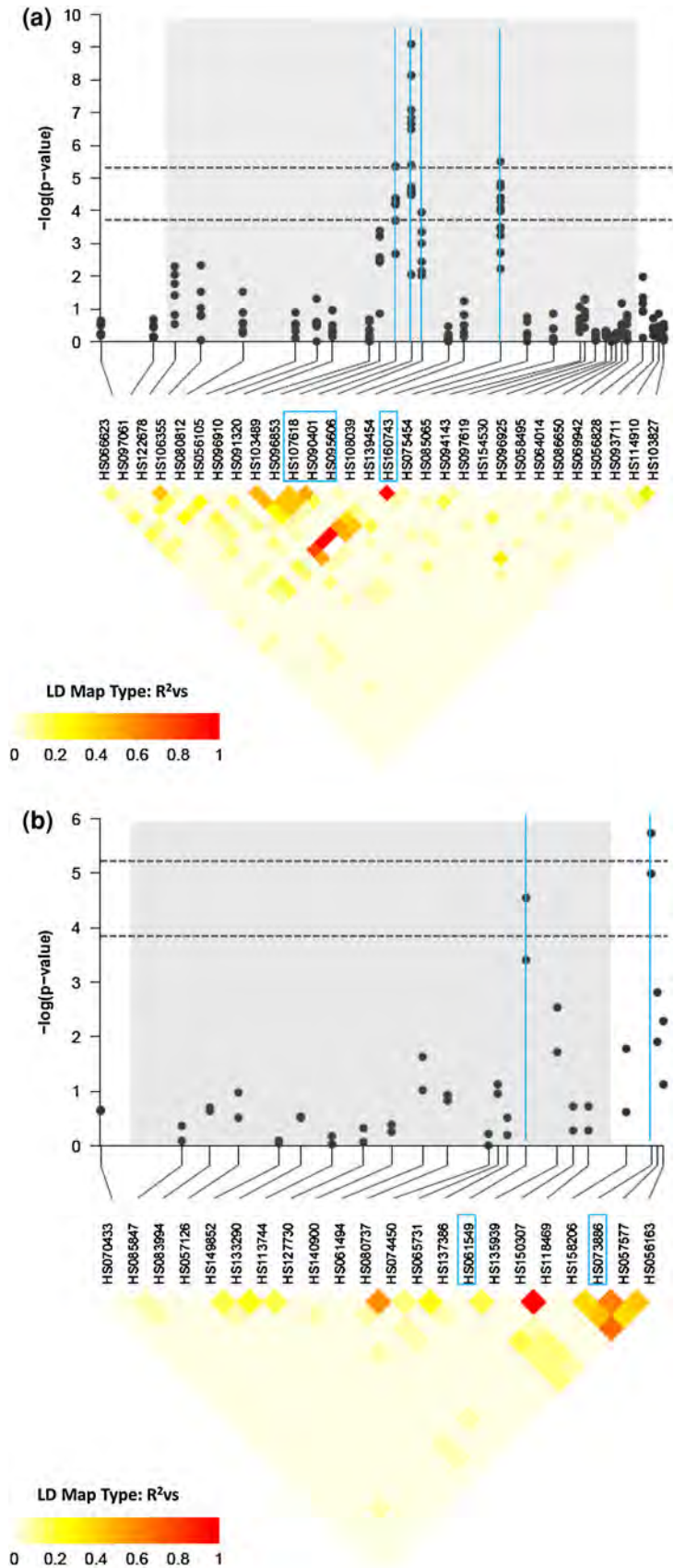
The third group (g3) was constituted with a large proportion of the public maintainer lines (B-lines).

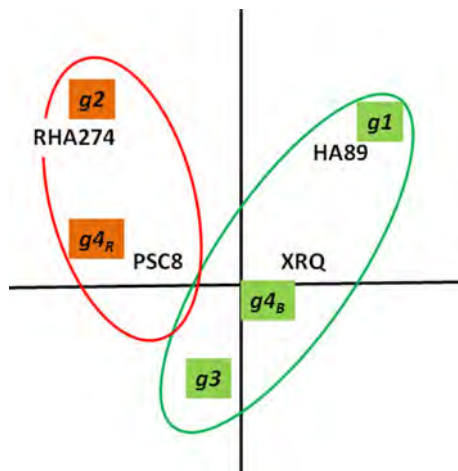
The other genotypes, both public or proprietary elite, maintainer or restorer lines were not assigned to a group by STRUCTURE, but for clarity will be designated group g4 (g4<sub>B</sub> and g4<sub>R</sub> for B lines and R lines, respectively).

The group g3 of B-lines showed a lower level of divergence from g4 compared to g1 and g2. In addition, g3 presented less diversity than the public R lines belonging to g4<sub>R</sub> as shown in Fig. 2b. This is in agreement with observations of Mandel et al. (2011), who used 86 of the set of public lines included in our panel as part of a collection of 433 cultivated accessions. Taking into account genotyping data from two SSR markers per linkage group, these authors observed less diversity among B lines than among the restorer lines (“INRA RHA” in their work). A large fraction of the restorer lines involved in breeding programs were derived from crosses between one of the main sources of fertility restoration genes (wild *H. annuus* carrying *Rf1*) like RHA274, and B lines (or their CMS counterpart). The founder effect of the line RHA274 could



**Fig. 7** LD heat map with corrected LD statistics ( $r_{vs}^2$ ). Only QTL regions (*shaded area*) and markers tested in association mapping with their  $p$  values are represented. Bonferroni (*lower*) and FDR thresholds are indicated by *dashed lines*. **a** LG10, over 10.7 cM **b** LG09, over 18 cM (color figure online)





**Fig. 8** Simplified representation of genetic groups according to Fig. 2b (color figure online)

explain why the R lines belonging to g4 have been found wide spread between RHA274, found close to g2, and more or less closely related to the g3 B line group.

In contrast, the large distance between HA89, which was derived from the Russian open pollinated population VNIIMK 8931, and the g3 group suggests that, in our panel, the set of B lines is divided into two subsets. One of these is more or less related to the HA89, and the other (g3) could be related to another particular founder effect which is not known. According to the pedigree information we have, this founder could have originated from Eastern European breeding programs, with either specific selections from Russian open pollinated varieties other than VNIIMK 8931, or introgressions from wild *Helianthus*, or exotic germplasm very different from g1 and g2.

Divergence between B and R elite gene pools has previously been mentioned (Gentzbittel et al. 1994). The fact that it did not appear as a major factor of our panel when using STRUCTURE is probably due to the presence of introgression lines. In addition, the three groups revealed by STRUCTURE or the PCA did not provide the best model for association detection. Indeed, the best models were obtained with the nearly statistically equivalent “ $K_{\text{ais}}$ ” or “ $K_{\text{ais}} + \text{Tester}$ ” models. Although these two models aimed at correcting different layers of stratification, the BIC values were quite similar. Furthermore, both of these models led to a good control of the rate of false positives and they detect the same most significant SNP. All but two of the detected associations were confirmed to be true positives, as they were located within regions detected through linkage mapping. It was not possible to check the validity of the two remnant associations, as the RIL parents were not polymorphic at these loci.

These results show that kinship probably accounted for most of the information provided by the structure of the

panel. They also validate our strategy of keeping a sub-optimal model such as “ $K_{\text{ais}}$ ” for association detection.

#### Linkage disequilibrium

Because the resolution of association mapping depends on the LD between genotyped markers and causative polymorphism, it was necessary to understand the nature of LD in our panel. LD decay is a good predictor of what is expected, given the present marker density. Standard measurements of LD have been described in the literature (Flint-Garcia et al. 2003). However, long range LD, which results from the presence of subpopulations with different allelic frequencies, can bias these estimates. We therefore used the modified statistics proposed by Mangin et al. (2011). This modified estimation, when compared to classical  $r^2$  statistics, lead to a considerably increased rate of LD decay by neglecting  $r^2$  between unlinked loci. When considering our design, this modification took kinship and the B/R differentiation into account. This is reminiscent of the “ $K_{\text{ais}} + \text{Tester}$ ” model which was found overall the best for association detection. Moreover, we investigated the LD decay in each breeding pool using the LD statistics, thereby correcting for the relatedness between lines in each group.

When considering all of the lines together, large variations of LD were observed between linkage groups. The extent of LD is influenced by many factors, such as recombination and selection (Gaut and Long 2003) through hitchhiking with selected loci (Maynard Smith and Haigh 1974; Mackay and Powell 2007). The breeding history of the parental lines of sunflower hybrids might have resulted in a high mean LD on LG08 and LG10 compared to other LGs. Indeed, the recessive branching gene is located on chromosome 10. Thus, positive or negative selection for this trait, depending on the breeding purpose (B or R lines, respectively), might have caused the extent of LD near this locus. Similarly, a cluster of resistance genes to downy mildew (including *PI2*, Bouzidi et al. 2002), mapped on LG08, has been an important target of selection, particularly for R lines during the 1975–1995 period.

Even using the corrected statistics, the overall LD in our panel (0.14 cM at 0.20 threshold) remained high. It extended over approximately 272 kb, with a 3.5-Gb genome and a genetic map of 1,800 cM. Comparisons with previous studies are difficult. LD extends from 50 to 250 kb in *Arabidopsis thaliana*, a self-pollinated species (Nordborg et al. 2005). In crops, elite lines can present a slower decay of LD (Rafalski 2002). For example, in maize, different studies have shown that LD persists over 1 kb for landraces (Tenaillon et al. 2001) to 500 kb for commercial elite inbred lines (Jung et al. 2004).

In sunflower, Kolkman et al. (2007) analyzed a set of ten elite inbred lines and two wild accessions and estimated LD decay for 30 loci. In this study, LD was found to extend over 5,500 bp for a  $r^2$  threshold of 0.32, suggesting that the threshold we used in this study would have led to a larger extent of LD. Fusari et al. (2008) assessed the LD over 28 candidate genes (<1 kb) in a panel of 19 elite inbred lines. They estimated that LD decays over 643 bp for an  $r^2$  of 0.64 in the entire set or 0.48 in one of the subpopulations identified by STRUCTURE. This result highlights the need to take into account the presence of subpopulations when estimating LD. These earlier studies assessed LD over short distances (<1 kb for the latter), whereas we investigated the LD in a broader sample that accounted for structure. A slow LD decay usually confers poor resolution for association mapping but requires fewer markers. In this study, given the LD estimated, 12,857 SNP would be required to cover the genome, but we only met half of this requirement.

On all LGs, the LD was found to decay more rapidly in the entire panel than within the B or R group. When considering all of the lines together, because selection pressures were different for B and R lines depending on trait, a different number of recombination events may have occurred for B lines and for R lines. Second, estimates of within-population rates of LD decay are subject to much larger standard errors than those based on whole populations, due to the smaller number of polymorphic sites (Ingvarsson 2005).

### Mapping results

Most significant associations were specific to a set of testers (83HR4gms/T1 or T2/T3). The consistency of mapping results between the testcross results is of great importance, especially when they are applied to breeding programs (Melchinger et al. 1998). Among the hypotheses that explain the lack of common QTLs between testers, the most common is that dominant alleles of tester lines, especially when they are elite lines (Hallauer and Miranda 1988; Austin et al. 2000), can exert masking effects. In the testing design we used, only dominant alleles from the panel are expected to be detected when combined with recessive alleles from the testers.

In the linkage mapping study, rather low correlations between environments were observed, while a relatively high congruency was found between the detected QTL. Due to this congruency, shorter interval supports for congruent QTL were found when running the QTL detection on the year or tester averages (data not shown). In contrast, in this association study, even when raw correlations between environments where high  $r^2$  (values ranging 0.42–0.93), only a few signals were repeated across

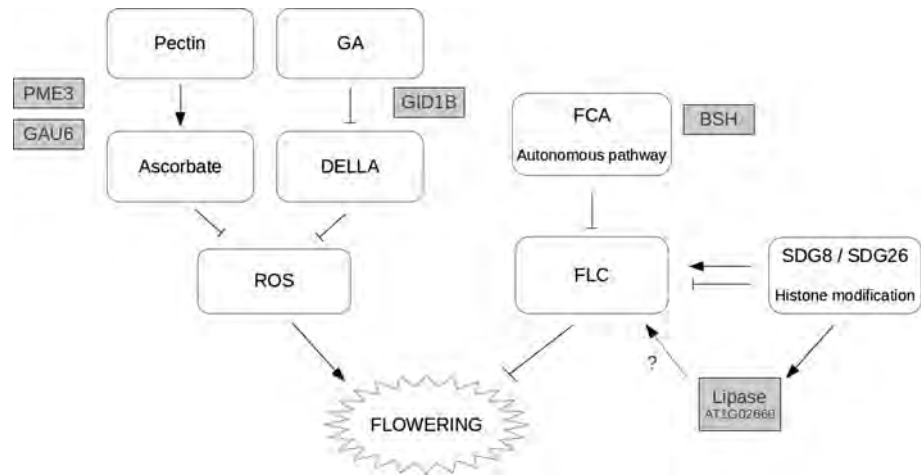
environments. In an attempt to explain this intriguing result, we first examined the impact of the panel relatedness to the high raw correlations. Using an approach similar to that of Mangin et al. (2011), we calculated a modified  $r^2$  using the kinship matrix  $K$  as a metrics. These modified  $r^2$  values were considerably lower than raw  $r^2$  values (0.00–0.58). This suggests that kinship—which also accounts for B-lines vs. R-lines—drove the common responses of the panel across the environments, whereas association peaks were the driving forces behind geno-type  $\times$  environment interactions. However, this does not explain why we found, in the linkage mapping study, lower correlations between environments together with more congruency of QTL. It should be pointed out that in the association study, only hybrids between B lines and R-type testers (83HR4gms or T2) or between R lines and B-type testers (T1 or T3) were evaluated (“unrelated” context). In contrast, each RIL involved in the linkage mapping study is a patchwork including B and R background. This patchwork was evaluated either in testcrosses with B-type and R-type testers, or per se, thus leading at least to local B  $\times$  B or R  $\times$  R combinations (“related” context). On another cross-pollinated species (*Medicago sativa*), it has been shown that part of the genetic variability expressed differs between “related” and “unrelated” contexts (Galais 1984). Moreover, hybrids (“unrelated” context) generally show better stability across environments, in the sense of Allard and Bradshaw (1964), than inbred lines (“related context”). This could explain why testcross data did not result in the same level of correlation between environments in the association study and in the linkage mapping study. Finally, we hypothesize that the covariance between environments in the linkage mapping study, which is still significant as shown in Fig. 1b, accounts for congruency between QTL.

All together these results indicate that despite its well-documented weakness, the linkage mapping approach is still relevant, because it is robust when relevant genetic variability exists between the parental lines of the RIL population.

Recently, several studies integrating association and linkage mapping have demonstrated the power and resolution of this approach to identify loci of interest. These joint-linkage association mapping methods are based on controlled crosses that provide equilibrated allelic frequencies (Myles et al. 2009). Such a design was developed in maize with a nested association mapping (NAM) population consisting of 25 RIL populations derived from crosses between 25 diverse lines and a tester line B73, which was used to dissect flowering times (Buckler et al. 2009).

Taking advantage of our combined linkage mapping analysis, we were able to determine whether the association

**Fig. 9** Candidate pathways involved in flowering time variation in sunflower. Genes carrying associated polymorphisms are indicated in gray boxes



signals we detected were false positives. A true association at a given locus implies that a QTL overlapping this locus should exist if population parents carry different alleles for that locus (Zhao et al. 2007). In this study, all of the SNP detected through the association mapping approach, and for which alleles differed between RIL parents, were located in LM-QTL regions. In the case of LG10 and LG15 (detected only through the suboptimal model “ $K_{ais}$ ”), we confirmed that the SNP detected were true positives whereas those detected on LG04, LG12 and LG17 could not be confirmed, as the RIL parents were not polymorphic at these loci.

On the contrary, one major QTL identified with linkage mapping on LG14 and also reported in the literature (Poormohammad Kiani et al. 2009) was not tagged through association mapping. This situation can occur when one of the alleles at the locus concerned is present at low frequency in the association panel. It can also occur when the structure of the panel is correlated with polymorphism at this locus, thus inducing a “false negative” (Famoso et al. 2011). The 106 SNP tested in association mapping were mapped within the 20-cM support interval of this LG14 QTL. We observed large blocks of LD in this region when no structure correction was applied, thus helping to confirm the hypothesis of a false negative in the association approach.

According to the genetic profile of the two parental lines XRQ and PSC8 (Fig. 8), the RIL population enabled the detection of differences resulting from their respective derivations:  $g1 \times g3$  for XRQ, and  $g2 \times g3$  for PSC8, which could make it possible to identify loci differentiating  $g1$  and  $g2$ . In contrast, in the association mapping design, kinship accounted for breeding history and it may have precluded the detection of association peaks on LG14. This also suggests that the QTL located on LG14 is an important feature distinguishing B-type and R-type sunflower lines for flowering time.

The polygenic pattern of inheritance of the flowering trait in sunflower has been reported in the literature (Leon et al. 2000). Taking into consideration the overlap between association and linkage mapping results, eight regions involved in the inheritance of flowering time were identified in this study. Several of these are in good agreement with other linkage mapping results from studies concerning the same trait.

In these eight regions, five potential candidates identified through association peaks appear to be functionally related to flowering time in other species (Fig. 9). The peptide predicted from the sunflower sequence HaT131016684 and corresponding to the associated marker HS107108 (localized on LG05) is homologous to the Arabidopsis BSH protein. Interestingly, BSH interacts with SWI3B to form a complex with the regulator of flowering time FCA (Sarnowski et al. 2002). The epigenetic control of FLC via H3K36 histone modification is controlled by SDG8 and SDG26 in Arabidopsis and could also be involved in sunflower. In fact, in their study, Xu et al. (2008) identified in both SDG8 and SDG26 mutant back-grounds a modification of expression of a lipase homologous to a gene associated with flowering time variation in our study: HaT131002875. More strikingly, HaT131048245, which carries the marker HS061549 on LG10, is homologous to Arabidopsis GID1B. This gibberellin receptor regulates DELLA proteins by targeting the proteins to the proteasome. It is well known that the gibberellin/DELLA pathway controls flowering time in Arabidopsis (Sun 2010). Several studies have confirmed that this pathway is prone to genetic variation, which explains flowering time variations in crops such as maize (Andersen et al. 2005; Thornsberry et al. 2001) and, most likely, canola (Raman et al. 2012). However, due to differences in genetic material, Blackman et al. (2011) were unable to confirm the presence of this pathway in sunflower. DELLA repressed ROS accumulation by increasing of the transcription of

genes encoding ROS scavenging enzymes (Achard et al. 2008). Interestingly, two other genes that regulate ROS accumulation via the ascorbate pathway in *Oncidium* (Shen et al. 2009) are carrying confirmed SNP: HaT131001758 and HaT131024508, which are homologous to GAUT6 and PME3, respectively. The functional analysis of genes associated with flowering time should still be considered with extreme caution. However, not surprisingly, different pathways have been identified in which ROS may play a role and further analysis is needed to resolve this issue.

## Conclusion

This study has shown, for observations of flowering time in cultivated sunflower across several genetic and agronomic environments, a large similarity between association peaks detected in a wide panel and QTL mapped in a RIL population. In the linkage disequilibrium mapping approach, comparison of models demonstrated that the kinship provided the best fit. However, both due to the genetic design which aimed to analyze other agronomic traits and to the breeding history, the detection of associations using this model did not identify a QTL documented in the literature and confirmed in this study. The results have provided novel information on whole genome linkage disequilibrium in cultivated sunflower, including a possible hitchhiking effect on two linkage groups carrying loci under strong selection during the breeding process.

**Acknowledgments** We would like to thank M.C. Boniface and D. Varès (INRA Toulouse), H. Bony, G. Joubert, F. Serre, S. Roche and J. Philippon (INRA Clermont-Ferrand), Th. André (SOLTIS), S. Châtre (RAGT), P. George and M. Barthes (BIOGEMMA) and colleagues from SYNGENTA Seeds for their involvement in sunflower trial management. This work benefited from the GENOPLANTE program “HP1” (2001–2004), the “SUNYFUEL” project, financially supported by the French National Research Agency (2008–2011), and the “OLEOSOL” project (2009–2012) with the financial support from the Midi Pyrénées Region, the European Fund for Regional Development (EFRD), and the French Fund for Competitiveness Clusters (FUI).

## References

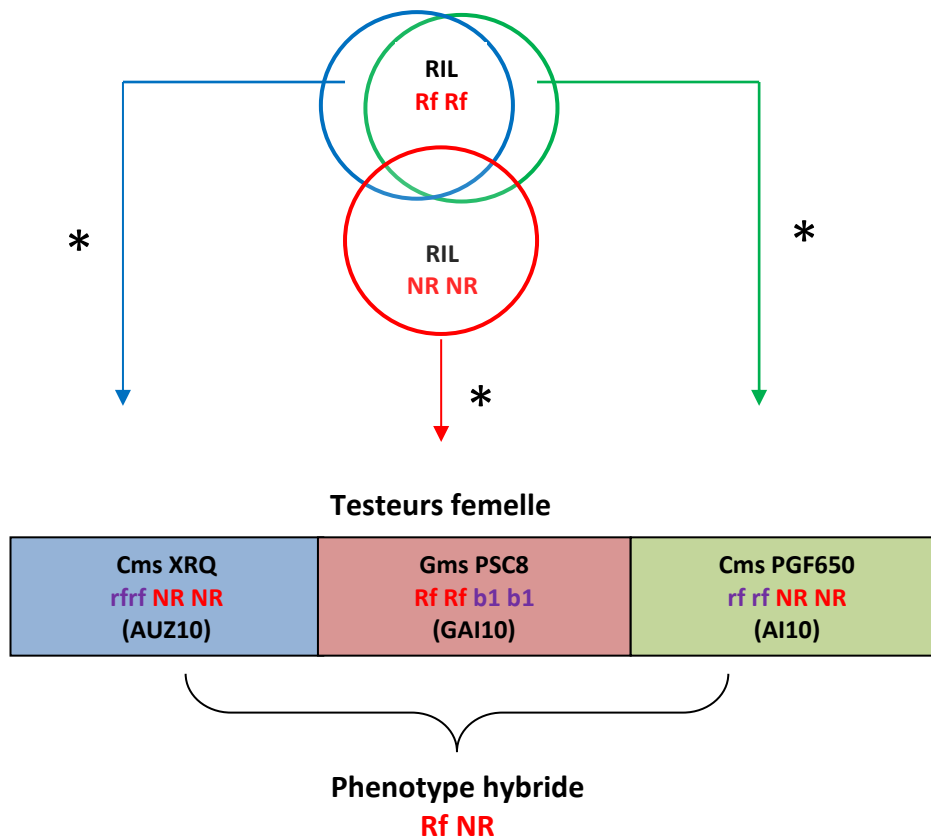
- Achard P, Renou JP, Berthomé R, Harberd NP, Genschik P (2008) Plant DELLAs restrain growth and promote survival of adversity by reducing the levels of reactive oxygen species. *Curr Biol* 18:656–660
- Allard RW, Bradshaw AD (1964) Implications of genotype-environmental interactions in applied plant breeding. *Crop Sci* 4:503–508
- Andersen JR, Schrag T, Melchinger AE, Zein I, Lübberstedt T (2005) Validation of Dwarf8 polymorphisms associated with flowering time in elite European inbred lines of maize (*Zea mays* L.). *Theor Appl Genet* 111:206–217
- Aranzana MJ, Kim S, Zhao K, Bakker E, Horton M et al (2005) Genome-wide association mapping in *Arabidopsis* identifies previously known flowering time and pathogen resistance genes. *PLoS Genet* 1:e60
- Arcade A, Labourdette A, Falque M, Mangin B, Chardon F, Charcosset A, Joets J (2004) BioMercator: integrating genetic maps and QTL towards discovery of candidate genes. *Bioinformatics* 20:2324–2326
- Atwell S, Huang YS, Vilhjalmsdottir BJ, Willems G, Horton M, Li Y et al (2010) Genome-wide association study of 107 traits in *Arabidopsis thaliana* inbred lines. *Nature* 465:627–631
- Austin DF, Lee M, Veldboom LR, Hallauer AR (2000) Genetic mapping in maize hybrid progeny across testers and generations: grain yield and grain moisture. *Crop Sci* 40:30–39
- Baack EJ, Sapir Y, Chapman MA, Burke JM, Rieseberg LH (2008) Selection on domestication traits and QTLs in crop-wild sunflower hybrids. *Mol Ecol* 17:666–677
- Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Stat Soc B Met* 57:289–300
- Bert PF, Jouan I, Tourvielle de Labrouhe D, Serre F, Philippon J, Nicolas P, Vear F (2003) Comparative genetic analysis of quantitative traits in sunflower (*Helianthus annuus* L.). 3. Characterisation of QTL involved in developmental and agronomic traits. *Theor Appl Genet* 107:181–189
- Blackman BK, Rasmussen DA, Strasburg JL, Raduski AR, Burke JM, Knapp SJ, Michaels SD, Rieseberg LH (2011) Contributions of flowering time genes to sunflower domestication and improvement. *Genetics* 187:271–287
- Bouzidi MF, Badaoui S, Cambon F, Vear F, De Labrouche DT, Nicolas P, Mouzeyar S (2002) Molecular analysis of a major locus for resistance to downy mildew in sunflower with specific PCR-based markers. *Theor Appl Genet* 104:600–952
- Bowers JE, Bachlava E, Brunick RL, Rieseberg LH, Knapp SJ, Burke JM (2012) Development of a 10,000 locus genetic map of the sunflower genome based on multiple crosses. *Genes Genomes Genetics* 2:721–729
- Brachi B, Faure N, Horton M, Flahauw E, Vazquez A et al (2010) Linkage and association mapping of *Arabidopsis thaliana* flowering time in nature. *PLoS Genet* 6:e1000940
- Breseghele F, Sorrells ME (2006) Association analysis as a strategy for improvement of quantitative traits in plants. *Crop Sci* 46:1323–1330
- Buckler ES, Holland JB, Bradbury PJ, Acharya CB, Brown PJ et al (2009) The genetic architecture of maize flowering time. *Science* 325:714–718
- Burke JM, Tang S, Knapp SJ, Rieseberg LH (2002) Genetic analysis of sunflower domestication. *Genetics* 161:1257–1267
- Butler DG, Cullis BR, Gilmour AR, Gogel BJ (2007) ASReml-R reference manual. The State of Queensland, Department of Primary Industries and Fisheries, Brisbane
- Charcosset A, Mangin B, Moreau L, Combes L, Jourjon MF (2000) Heterosis in maize investigated using connected RIL populations. In: Quantitative genetics and breeding methods: the way ahead. INRA, Paris, pp 89–98
- Coque M, Mesnildrey S, Romestant M, et al (2008) Sunflower lines core collections for association studies and phenomics. In: Proceedings ASTA Conference, Cordoba
- Crouzillat D, De la Canal L, Perrault A, Ledoigt G, Vear F, Serieys H (1991) Cytoplasmic male sterility in sunflower: comparison of molecular biology and genetic studies. *Plant Mol Biol* 16:415–426
- Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol Ecol* 14:2611–2620

- Famoso AN, Zhao K, Clark RT, Tung C, Wright MH, Kochian LV, McCouch SR (2011) Genetic architecture of aluminum tolerance in rice (*Oryza sativa*) determined through genome-wide association analysis and QTL mapping. *PLoS Genet* 7:e1002221
- Flint-Garcia SA, Thornsberry JM, Buckler ES (2003) Structure of linkage disequilibrium in plants. *Annu Rev Plant Biol* 54:357–374
- Flint-Garcia SA, Thuillet A-C, Yu J, Pressoir G, Romero SM, Mitchell SE, Doebley J, Kresovich S, Goodman MM, Buckler ES (2005) Maize association population: a high-resolution platform for quantitative trait locus dissection. *Plant J* 44:1054–1064
- Fusari CM, Lia VV, Hopp HE, Heinz RA, Paniago NB (2008) Identification of single nucleotide polymorphisms and analysis of linkage disequilibrium in sunflower elite inbred lines using the candidate gene approach. *BMC Plant Biol* 8:7
- Fusari CM, Rienzo JA, Troglia C (2012) Association mapping in sunflower for *Sclerotinia* head rot resistance. *BMC Plant Biol* 12:93
- Gallais A (1984) An analysis of heterosis vs. inbreeding effects with an autotetraploid cross-fertilized plant *Medicago sativa* L. *Genetics* 106:123–137
- Gaut BS, Long AD (2003) The lowdown on linkage disequilibrium. *Plant Cell* 15:1502–1506
- Gentzbittel L, Zhang YX, Vear F, Griveau B, Nicolas P (1994) RFLP studies of genetic relationships among inbred lines of the cultivated sunflower, *Helianthus annuus* L.: evidence for distinct restorer and maintainer germplasm pools. *Theor Appl Genet* 89:419–425
- Hallauer AR, Miranda JB (1988) Quantitative genetics in maize breeding. Iowa State Univ Press, Ames
- Hill WG, Weir BS (1988) Variances and covariances of squared linkage disequilibria in finite populations. *Theor Popul Biol* 33:54–78
- Holland JB (2004) Implementation of molecular markers for quantitative traits in breeding programs—challenges and opportunities. In: Fischer T, Turner N, Angus J, McIntyre L, Robertson M, Borrell A, Lloyd D (eds) New directions for a diverse planet: proceedings for the 4th international crop science congress. Brisbane, Australia
- Horne EC, Kumpatla SP, Patterson KA, Gupta M, Thompson SA (2004) Improved high-throughput sunflower and cotton genomic DNA extraction and PCR fidelity. *Plant Mol Biol Rep* 22:83a–83i
- Ingarvarsson PK (2005) Nucleotide polymorphism and linkage disequilibrium within and among natural populations of European aspen (*Populus tremula* L., Salicaceae). *Genetics* 169:945–953
- Jourjon MF, Jasson S, Marcel J, Ngom B, Mangin B (2005) MCQTL: multi-allelic QTL mapping in multi-cross design. *Bioinformatics* 21:128–130
- Jung M, Ching A, Bhatramakki D, Dolan M, Tingey S et al (2004) Linkage disequilibrium and sequence diversity in a 500-kbp region around the *adh1* locus in elite maize germplasm. *Theor Appl Genet* 109:681–689
- Kane NC, Gill N, King MG, Bowers JE, Berges H et al (2011) Progress towards a reference genome for sunflower. *Botany* 89:429–437
- Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, Eskin E (2008) Efficient control of population structure in model organism association mapping. *Genetics* 178:1709–1723
- Kolkman JM, Berry ST, Leon AJ, Slabaugh MB, Tang S, Gao W et al (2007) Single nucleotide polymorphisms and linkage disequilibrium in sunflower. *Genetics* 177:457–468
- Lander ES, Botstein DB (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* 121:185–199
- Leon AJ, Andrade FH, Lee M (2000) Genetic mapping of factors affecting quantitative variation for flowering in sunflower. *Crop Sci* 40:404–407
- Leon AJ, Lee M, Andrade FH (2001) Quantitative trait loci for growing degree days to flowering and photoperiod response in Sunflower (*Helianthus annuus* L.). *Theor Appl Genet* 102:497–503
- Luna A, Nicodemus KK (2007) snp.plotter: an R-based SNP/haplotype association and linkage disequilibrium plotting package. *Bioinformatics* 23:774–776
- Mackay TFC (2001) The genetic architecture of quantitative traits. *Annu Rev Genet* 35:303–339
- Mackay I, Powell W (2007) Methods for linkage disequilibrium mapping in crops. *Trends Plant Sci* 12:57–63
- Maenhout S, De Baets B, Haesaert G (2009) Marker-based estimation of the coefficient of coancestry in hybrid breeding programmes. *Theor Appl Genet* 118:1181–1192
- Mandel JR, Dechaine JM, Marek LF, Burke JM (2011) Genetic diversity and population structure in cultivated sunflower and a comparison to its wild progenitor, *Helianthus annuus* L. *Theor Appl Genet* 123:693–704
- Mangin B, Siberchicot A, Nicolas S, Doligez A, This P et al (2011) Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. *Heredity* 108:285–291
- Maynard Smith J, Haigh J (1974) The hitch-hiking effect of a favorable gene. *Genet Res* 23:23–35
- Melchinger AE, Utz HF, Schon CC (1998) Quantitative trait locus (QTL) mapping using different testers and independent population samples in maize reveals low power of QTL detection and large bias in estimates of QTL effects. *Genetics* 149:383–403
- Mestries E, Gentzbittel L, Labrouhe DT, Nicolas P, Vear F, Am S (1998) Analyses of quantitative trait loci associated with resistance to *Sclerotinia sclerotiorum* in sunflowers (*Helianthus annuus* L.) using molecular markers. *Mol Breed* 4:215–226
- Mir RR, Kumar N, Jaiswal V, Girdharwal N, Prasad M, Balyan HS, Gupta PK (2012) Genetic dissection of grain weight in bread wheat through quantitative trait locus interval and association mapping. *Mol Breed* 29:963–972
- Mokrani L, Gentzbittel L, Azanza F, Fitamant L, Al-Chaarani G, Sarrafi A (2002) Mapping and analysis of quantitative trait loci for grain oil and agronomic traits using AFLP and SSR in sunflower (*Helianthus annuus* L.). *Theor Appl Genet* 106:149–156
- Myles S, Peiffer J, Brown PJ, Ersoz ES, Zhang Z, Costich DE, Buckler ES (2009) Association mapping: critical considerations shift from genotyping to experimental design. *Plant Cell* 21:2194–2220
- Nordborg M, Weigel D (2008) Next-generation genetics in plants. *Nature* 456:720–723
- Nordborg M, Hu TT, Ishino Y, Jhaveri J, Toomajian C, Zheng H, Bakker E, Calabrese P, Gladstone J, Goyal R, Jakobsson M, Kim S, Morozov Y, Padhukasahasram B, Plagnol V, Rosenberg NA, Shah C, Wall J, Wang J, Zhao K, Kalbfleisch T, Schultz V, Kreitman M, Bergelson J (2005) The pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biol* 3:e196
- Patterson N, Price AL, Reich D (2006) Population structure and Eigen analysis. *PLoS Genet* 2:e190
- Poormohammad Kiani S, Maury P, Nouri L, Ykhlef N, Grieu P, Sarrafi A (2009) QTL analysis of yield-related traits in sunflower under different water treatments. *Plant Breeding* 128:363–373
- Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* 155:945–959
- R Development Core Team (2012) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. ISBN 3-900051-07-0

- Rafalski JA (2002) Novel genetic mapping tools in plants: SNPs and LD-based approaches. *Plant Sci* 162:329–333
- Raman H, Raman R, Eckermann P et al (2012) Genetic and physical mapping of flowering time loci in canola (*Brassica napus* L.). *Theor Appl Genet* 126:119–132
- Rieseberg LH, Willis JH (2007) Plant speciation. *Science* 317:910–914
- Sarnowski TJ, Świesz S, Pawlikowska K, Kaczanowski S, Jerzmanowski A (2002) AtSWI3B, an Arabidopsis homolog of SWI3, a core subunit of yeast Swi/Snf chromatin remodeling complex, interacts with FCA, a regulator of flowering time. *Nucl Acids Res* 30:3412–3421
- Shen CH, Krishnamurthy R, Yeh KW (2009) Decreased L-ascorbate content mediating bolting is mainly regulated by the galacturonate pathway in *Oncidium*. *Plant Cell Physiol* 50:935–946
- Sun TP (2010) Gibberellin-GID1-DELLA: a pivotal regulatory module for plant growth and development. *Plant Physiol* 154:567–570
- Sun G, Zhu C, Kramer MH, Yang SS, Song W, Piepho HP, Yu J (2010) Variation explained in mixed-model association mapping. *Heredity* 105:333–340
- Tenaillon MI, Sawkins MC, Long AD, Gaut RL, Doebley JF et al (2001) Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). *Proc Natl Acad Sci USA* 98:9161–9166
- Thornsberry JM, Goodman MM, Doebley J et al (2001) Dwarf8 polymorphisms associate with variation in flowering time. *Nature genet* 28:286–289
- Tian F et al (2011) Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nat Genet* 43:159–162
- Vear F, Serre F, Jouan-Dufournel, Bert I, Roche PF, Walser SP, de Labrouhe DT, Vincourt P (2008) Inheritance of quantitative resistance to downy mildew (*Plasmopara halstedii*) in sunflower (*Helianthus annuus* L.). *Euphytica* 164:561–570
- Vincourt P, As Sadi F, Bordat A, Langlade N, Gouzy J, Pouilly N, Lippi Y, Serre F, Godiard L, Tourvieille de Labrouhe D, Vear F (2012) Consensus mapping of major resistance genes and independent QTL for quantitative resistance to sunflower downy mildew. *Theor Appl Genet* 5:909–920
- Wills DM, Burke JM (2007) Quantitative trait locus analysis of the early domestication of sunflower. *Genetics* 176:2589–2599
- Wright S (1951) The genetic structure of populations. *Ann Eugen* 15:323–354
- Xu L, Zhao Z, Dong A et al (2008) Di- and tri- but not monomethylation on histone H3 lysine 36 marks active transcription of genes involved in flowering time regulation and other processes in *Arabidopsis thaliana*. *Mol Cell Biol* 28:1348–1360
- Yan J, Warburton ML, Crouch J (2011) Association mapping for enhancing maize (*Zea mays* L.) genetic improvement. *Crop Sci* 51:433–449
- Yu JM, Pressoir G, Briggs WH, Bi IV, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet* 38:203–208
- Zhao K, Aranzana MJ, Kim S, Lister C, Shindo C, Tang C, Toomajian C, Zheng H, Dean C, Marjoram P, Nordborg M (2007) An Arabidopsis example of association mapping in structured samples. *PLoS Genet* 3:e4

- Rafalski JA (2002) Novel genetic mapping tools in plants: SNPs and LD-based approaches. *Plant Sci* 162:329–333
- Raman H, Raman R, Eckermann P et al (2012) Genetic and physical mapping of flowering time loci in canola (*Brassica napus* L.). *Theor Appl Genet* 126:119–132
- Rieseberg LH, Willis JH (2007) Plant speciation. *Science* 317:910–914
- Sarnowski TJ, Świesz S, Pawlikowska K, Kaczanowski S, Jerzmanowski A (2002) AtSWI3B, an Arabidopsis homolog of SWI3, a core subunit of yeast Swi/Snf chromatin remodeling complex, interacts with FCA, a regulator of flowering time. *Nucl Acids Res* 30:3412–3421
- Shen CH, Krishnamurthy R, Yeh KW (2009) Decreased L-ascorbate content mediating bolting is mainly regulated by the galacturonate pathway in *Oncidium*. *Plant Cell Physiol* 50:935–946
- Sun TP (2010) Gibberellin-GID1-DELLA: a pivotal regulatory module for plant growth and development. *Plant Physiol* 154:567–570
- Sun G, Zhu C, Kramer MH, Yang SS, Song W, Piepho HP, Yu J (2010) Variation explained in mixed-model association mapping. *Heredity* 105:333–340
- Tenaillon MI, Sawkins MC, Long AD, Gaut RL, Doebley JF et al (2001) Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). *Proc Natl Acad Sci USA* 98:9161–9166
- Thornsberry JM, Goodman MM, Doebley J et al (2001) Dwarf8 polymorphisms associate with variation in flowering time. *Nature genet* 28:286–289
- Tian F et al (2011) Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nat Genet* 43:159–162
- Vear F, Serre F, Jouan-Dufournel, Bert I, Roche PF, Walser SP, de Labrouhe DT, Vincourt P (2008) Inheritance of quantitative resistance to downy mildew (*Plasmopara halstedii*) in sunflower (*Helianthus annuus* L.). *Euphytica* 164:561–570
- Vincourt P, As Sadi F, Bordat A, Langlade N, Gouzy J, Pouilly N, Lippi Y, Serre F, Godiard L, Tourvieille de Labrouhe D, Vear F (2012) Consensus mapping of major resistance genes and independent QTL for quantitative resistance to sunflower downy mildew. *Theor Appl Genet* 5:909–920
- Wills DM, Burke JM (2007) Quantitative trait locus analysis of the early domestication of sunflower. *Genetics* 176:2589–2599
- Wright S (1951) The genetic structure of populations. *Ann Eugen* 15:323–354
- Xu L, Zhao Z, Dong A et al (2008) Di- and tri- but not monomethylation on histone H3 lysine 36 marks active transcription of genes involved in flowering time regulation and other processes in *Arabidopsis thaliana*. *Mol Cell Biol* 28:1348–1360
- Yan J, Warburton ML, Crouch J (2011) Association mapping for enhancing maize (*Zea mays* L.) genetic improvement. *Crop Sci* 51:433–449
- Yu JM, Pressoir G, Briggs WH, Bi IV, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet* 38:203–208
- Zhao K, Aranzana MJ, Kim S, Lister C, Shindo C, Tang C, Toomajian C, Zheng H, Dean C, Marjoram P, Nordborg M (2007) An Arabidopsis example of association mapping in structured samples. *PLoS Genet* 3:e4





**Figure V.I : Description du dispositif de croisements.**

Cms : cytoplasme male stérile, Gms : genique male sterile, rf : allèle récessif de restauration de la fertilité, Rf : allèle dominant de restauration de la fertilité, NR: allèle dominant de non ramification, b1 : allèle récessif de ramification

Environnement	Lieu	Année	Condition	Testeur pour lignées NR	Testeur pour lignées R	Nombre d'hybrides évalués
AUZ10_I	Auzeville	2010	irrigué	-	Cms XRQ	99
AUZ10_NI	Auzeville	2010	non irrigué	-	Cms XRQ	99
AI10_I	Aigrefeuille	2010	irrigué	-	CmsPGF650	150
AI10_NI	Aigrefeuille	2010	non irrigué	-	CmsPGF650	150
GA10_I	Gaillac	2010	irrigué	PSC8RMgms	-	113
GA10_NI	Gaillac	2010	non irrigué	PSC8RMgms	-	113

**Table V.1 : Détails des 6 environnements sur lesquels ont été évalués les RILS pour l'ensemble des caractères agronomiques.** Chaque lieu correspond à un testeur différent selon si la lignée est restauratrice de fertilité R ou non ramifiée (NR). Lieux : Auzeville (31) Gaillac (81) Aigrefeuille (17)

### V.3 Etude sur les autres caractères agronomiques

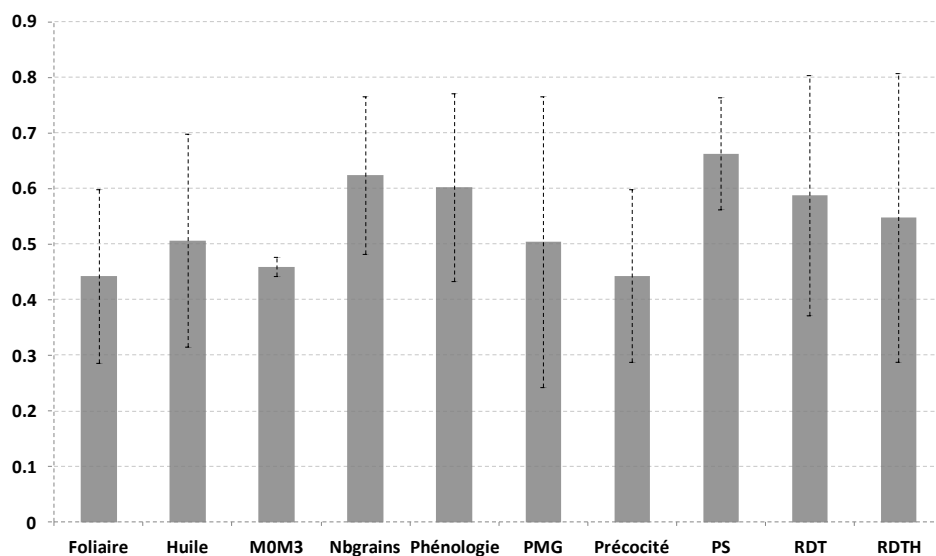
#### V.3.1 Matériels et méthodes

La même approche a été étendue aux autres caractères agronomiques mais en utilisant la carte génétique développée par Biogemma sur la population INEDI à partir de marqueurs non publics pour les détections de QTL. Cette carte a été réalisée à partir d'un sous ensemble de 214 RILs sur les 273 utilisées précédemment et 3985 marqueurs dont 3758 SNP. Le dispositif d'expérimentation pour la population biparentale est constitué de 3 lieux à deux conditions (irrigué ou non) et un testeur différent par lieu de manière à évaluer au champ des hybrides agronomiques, non ramifiés et si possibles mâles fertiles (Table V.1). 99 RILs restauratrices de fertilité mâle ont été croisées avec la lignée CMS XRQ sur les environnements AUZ10\_I et AUZ10\_NI. 150 RILs restauratrices (dont 85 communes avec le lieu précédent) ont été croisées avec la lignée CMS-PGF650 sur les environnements AI10\_I et AI10\_NI. Enfin, 113 RILs non ramifiées (l'absence de ramification est un caractère dominant par rapport à la ramification apicale classiquement déployée chez les lignées restauratrices de fertilité mâle) ont été croisées avec la version « mâle stérile génique » (gms) et résistante au mildiou (RM) de la lignée ramifiée PSC8, restauratrice de fertilité dans sa version initiale, et testées sur les environnements GAI10\_I et GAI10\_NI. Au final, un certain nombre de RIL sont communes à chaque paire de lieu mais seules 34 RILs ont pu être évaluée sur les 3 lieux. Le dispositif est résumé dans la figure V.1

A l'exception du diamètre au collet, les caractères mesurés sur la population sont identiques à ceux mesurés sur le panel : rendement (grains et huile), précocité de récolte (H2O), PMG, hauteur, phénologie, LAI, Indice foliaire. Le diamètre au collet est un paramètre de vigueur, assez bien corrélé au rendement chez le tournesol. Le dispositif des essais est similaire à celui utilisé pour le panel, mais les 2 modalités (sec/irrigué) sont présentes dans chacun des lieux. Chaque essai est divisé en unités de 24 à 36 hybrides, répétés en deux blocs et comprenant chacun les 4 témoins : INEDI, MELODY, PEGASOL, TEKNY. Le dispositif à Auzeville est cependant plus complet (de type lattice avec 3 répétitions) car voué à mesurer d'autres caractères plus finement dans le cadre d'un projet différent. Toutes les données phénotypiques ont été traitées en suivant la procédure présentée dans le chapitre II. Les prédictions génétiques sous formes de BLUP ont été extraites du meilleur modèle statistique après avoir comparé plusieurs modèles dont deux modèles spatiaux. Les corrélations phénotypiques ont été calculées entre toutes les combinaisons caractère-environnement, appelées variables, à partir du coefficient de Pearson (logiciel R). Chaque variable a été traitée séparément pour la

Caractère	AI10_I	AI10_NI	AUZ10_I	AUZ10_NI	GA10_I	GAI10_NI
Diamètre collet			NS	NS		
hauteur	x	x	NS	NS	x	x
H2O	x	x	x	x	x	x
IF40					x	x
IFF1					x	x
IFF20					x	x
LAD			x	x	x	x
LAI_F1			x	x	x	x
LAI_F20			x	NS	x	x
LAI_F40			x	x	x	x
Pente_if			NS	x		
Senescevol					x	x
Huile	x	x			x	x
F1	x	x	x	x	x	x
F1M0			x	x	x	NS
M0			x	x	x	x
MOM3			x	x	x	x
M3			x	x	x	x
Nbgrains			x	x	x	x
PMG			x	NS	x	x
PS			x	x		
RDT	NS	x	x	x	x	x
RDTH	NS	x			x	x

**Table V.2 : Résumé des caractères mesurés sur chaque environnement et de la significativité de leur effet génotypique (NS : mesuré et non significatif, X : mesuré et significatif)**



**Figure V.2 Héritabilités issues du modèle naïf par type de caractères (« Foliaire » regroupe les notes de LAI et d'indice foliaire, « Phénologie » regroupe les stades F1, M0, M3 et PS signifie poids spécifique)**

détection de QTL (suivant le même modèle que celui décrit dans l'article) en utilisant le logiciel MCQTL et la méthode Iterative Composite Interval Mapping. Le seuil de rejet du test a été calculé pour un risque de 5 % en effectuant 1000 permutations par variable. L'intervalle support de LOD utilisé est de 2 et la significativité des QTL a été évaluée grâce au test MCQTL ( $-\log(p\text{-value (Ftest)})$ ). Enfin, les positions des associations détectées dans la partie 2 ont été comparées aux QTL identifiés et la significativité de l'ensemble des colocalisations a été testée suivant Tian *et al.*, 2011 (cf. détails dans l'article). Tous les QTL ont été représentés sur une carte consensus appartenant à Biogemma et comprenant 8218 marqueurs à l'aide du logiciel BioMercator v4 (Arcade *et al.*, 2004).

### V.3.2 Résultats et discussion

#### V.3.2.1 Analyses phénotypiques

Sur l'ensemble des environnements, 86 variables phénotypiques ont été mesurées sur la population INEDI, 10 d'entre elles ont une variance génétique non significative ( $Z\text{ratio} < 2$ ) dont le rendement grain et huile à AI10\_I et le PMG à AUZ10\_NI (Table V.2). Sur le lieu AI10, les RILs croisées avec le testeur PGF650 présentent beaucoup de centres stériles, surtout en condition irriguée. Environ 15 % des parcelles ont été enlevées en raison de problèmes de peuplement, de plus, peu de caractères ont pu être mesurés sur ce lieu. L'héritabilité moyenne sur l'ensemble des caractères (0.48) y est également plus faible que pour AUZ10 (0.51) et GA10 (0.56).

Sur le lieu AUZ10, le stress hydrique a un effet opposé à celui attendu avec un rendement passant de 23 quintaux en condition irriguée à 28 quintaux en condition sèche. Certains paramètres liés au sol, aux précédents, ou encore au désherbage, pourraient expliquer ce phénomène. Quant à GAI10, on observe une augmentation de l'ordre de 6 quintaux du rendement avec irrigation. Les héritabilités moyennes par caractère (calculées à partir du modèle naïf) s'étendent entre 0.44 et 0.66 par lieu et par variable (Figure V.2) et sont en général plus faibles que celles enregistrées sur le panel d'association sauf pour le rendement et rendement en huile où certains lieux ont atteint des valeurs élevées. (maximum à 0.83 au lieu de 0.71 pour le panel). Cependant, la comparaison de 3 lieux pour la population avec 11 lieux pour le panel n'est pas suffisamment équilibrée pour conclure sur la différence de variabilité entre ces deux types de matériel, d'autant que les variances génétiques sont proches. En comparant les valeurs d'AIC, 79% des variables phénotypiques obtiennent un meilleur ajustement grâce à un modèle spatial. Les corrélations moyennes sur l'ensemble des

	AI10_I	AI10_NI	AUZ10_I	AUZ10_NI	GA10_I	GA10_NI
AI10_I	-	117	72	72	61	61
AI10_NI	0.16	-	72	72	61	61
AUZ10_I	0.11	0.06	-	83	37	37
AUZ10_NI	0.05	-0.01	0.10	-	37	37
GA10_I	0.09	0.08	0.10	0.10	-	97
GA10_NI	0.14	0.13	0.17	0.09	0.20	-

**Table V. 3: Moyenne des corrélations phénotypiques entre les environnements. Les effectifs communs aux paires d'environnement sont indiqués en rouge au-dessus de la diagonale.**

caractères pour toutes les paires de lieux ont été calculées (Table V.3). Celles-ci sont très faibles et atteignent un maximum de 0.20 entre les 2 conditions irriguée et sèche du lieu GAI10. GAI10\_NI est davantage corrélé avec tous les autres lieux. La présence de testeurs différents entre les 3 lieux expliquent la faiblesse des corrélations mais même au sein d'un même lieu, il n'y a que peu de corrélations entre irrigué et sec. Ces corrélations rendent compte de très fortes interactions testeurs \* RIL \* environnement, et donc de la difficulté de comparer ces deux conditions, en utilisant par exemple des index de tolérance au stress. Chaque variable phénotypique mesuré dans chaque environnement a donc été traitée séparément pour la détection de QTL.

#### V.3.2.2 *Détection de QTL*

Au total, 61 QTL ont été détectés sur 13 des 17 chromosomes (Table V.4 et Figure V.3). Plusieurs caractères sont très polygéniques : la floraison est associée à dix QTL répartis sur six chromosomes, la variable Nbgrains à neuf QTL répartis sur sept chromosomes et la hauteur à neuf QTL répartis sur cinq chromosomes. Ces trois caractères figurent parmi les plus héritables, mais à l'inverse, le rendement grain, caractère héritable dans cinq environnements sur les six (en moyenne  $h^2=0.66$ ) ne permet d'identifier que trois QTL sur deux zones différentes. Le rendement est un caractère complexe avec probablement un déterminisme génétique polygénique qui nécessiterait peut être une résolution plus fine et davantage de diversité que celle présente dans la population de RILs. En effet, la génétique d'association nous avait permis de détecter des signaux significatifs sur six chromosomes différents pour le rendement grains. Pour la floraison, on détecte environ le même nombre de signaux avec le panel et la population biparentale. Les détections de QTL pour la floraison ont été réalisées soit, dans le cadre de l'article, avec une carte publique de 517 marqueurs et une population de 273 RIL, soit dans le contexte présent avec 3985 marqueurs et un sous ensemble de 214 RILs. Les résultats sont globalement similaires mais la carte publique ne permet pas d'identifier deux QTL sur les chromosomes 1 et 8 tandis que la carte privée n'identifie pas un QTL sur le chromosome 16 (cf. article). Les 2 dispositifs sont donc complémentaires et apportent pour l'un plus de densité de marquage et pour l'autre plus de puissance avec davantage d'individus.

Pour les autres caractères, le panel amène en général à l'identification de plus de zones d'intérêt différentes que la population. Un exemple concerne le caractère précocité de récolte (H2O) où le panel permet d'identifier de nombreux signaux sur 15 chromosomes alors qu'aucun QTL n'a pu être détecté pour la population. La variabilité de la distribution des



BLUP pour ce caractère est plus importante chez le panel que dans la population des RIL, et constitue probablement une bonne explication pour ce résultat.

Au contraire pour Nbgrains, le panel ne permet pas de détecter autant de régions que la population. Les variances ne semblent pas être à l'origine de cette différence. Par contre, il est possible que les gènes responsables de ce caractère se situent dans des régions fortement influencés par la structure et donc non détectables avec les méthodes de corrections des faux positifs prenant en compte la structure. Nous reviendrons sur cette hypothèse dans la partie « colocalisation ».

L'une des deux zones où est identifié un QTL de rendement est localisée sur le LG10. Dans une fenêtre très élargie d'environ 30 cM, tous les QTL liés à des caractères associés à la productivité colocalisent (RDT, RDTH, PMG, Nbgrains et M0M3) ainsi que la hauteur et l'IFF1. Cette fenêtre rassemble les QTL avec les variances expliquées les plus fortes ; ainsi on observe un QTL de rendement s'étendant sur 5 cM avec un  $R^2$  de 0.32. Cette région, dont les allèles favorables sont apportés par le parent XRQ, semble donc très intéressante mais elle est spécifique au lieu GAI10. On n'observe en réalité que très peu de colocalisations entre lieux différents, ce qui est cohérent avec la faiblesse des corrélations mentionnée ci-dessus. La majorité des QTL sont identifiés à partir du lieu GAI10. Pour AI10, seules 3 régions sont identifiées pour leur rôle probable dans la variation de caractères phénologiques et de la hauteur, dont une colocalise avec des QTL identifiés sur GAI10. Ce lieu n'avait par ailleurs que très peu de variables mesurées disponibles. Quant à AUZ10, il permet d'identifier des QTL spécifiques sur le chromosome 8 (floraison et LAI) et sur le chromosome 16 (PMG et PS). A part sur le chromosome 1 (où les intervalles sont très grands), les chromosomes 6 et 9 sont les seuls où l'on peut observer des colocalisations entre lieux.

La deuxième zone associée à la productivité est identifiée sur le chromosome 5 et réunit des QTL de LAD, RDT et RDTH sur des intervalles assez courts (1, 4 et 7 cM respectivement).

Les pourcentages de variances phénotypiques expliquées par les QTL s'étendent de 14 à 32%, ce qui est largement supérieur aux variances expliquées par les SNP associés (inférieures à 10%). Malgré une héritabilité forte pour certains caractères, la variance expliquée par un SNP peut-être faible car le marqueur n'est pas en DL complet avec le gène causal. Il apparaît donc important de densifier en marqueurs les zones intéressantes. Quant à la variance expliquée par les QTL, il a été démontré qu'elle est probablement surestimée (Beavis, 1998).



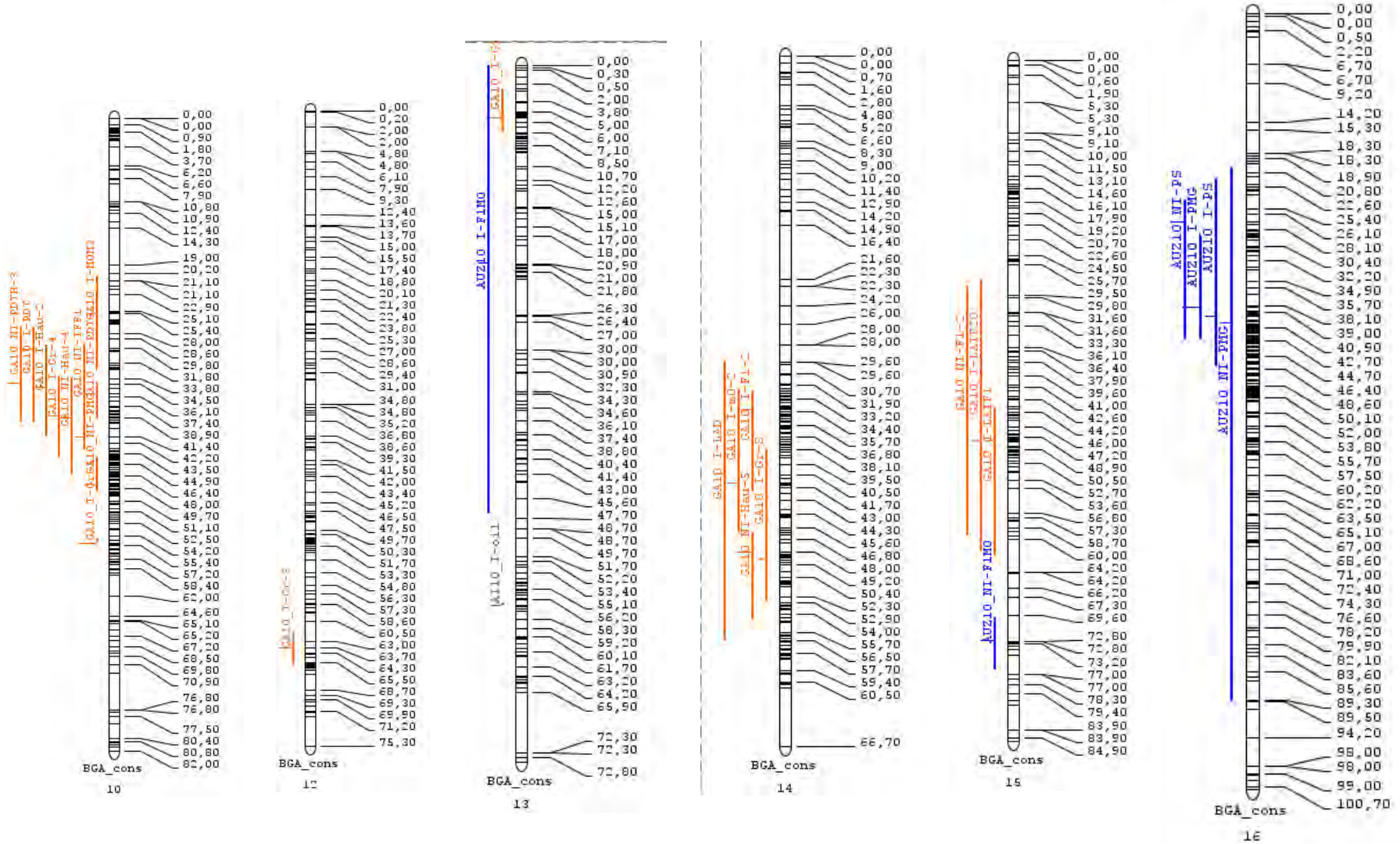


Figure V.3 (suite): Représentation des QTL sur la carte génétique à l'aide du logiciel BioMercator. Une couleur différente est utilisée par lieu.

### V.3.2.3 Colocalisations entre les associations et QTL bi parentaux

En considérant les SNP polymorphes entre les parents de la population de RIL, 55% des SNP identifiés par association sont présents dans des QTL mis en évidence grâce à la population biparentale. Ce taux est plus important pour les SNP détectés avec les deux modèles Kais et Kais + Testeur (64%) que ceux détectés qu'avec le modèle Kais (51%). Cette différence provient principalement des SNP identifiés sur les chromosomes 10 et 13. En effet, de grands blocs de DL ont été découverts sur ces chromosomes, probablement dû à la structure B/R (Cadic *et al.*, 2013), ce qui pourrait expliquer que ces SNP étaient probablement des faux positifs. Sur les 13 chromosomes pour lesquels des associations avaient été détectées, 5 ne présentent aucune colocalisation avec des QTL (LG03, 04, 11 et 17), il pourrait également s'agir de faux positifs ou d'allèles spécifiques au panel. Avec le modèle Kais + Testeur, la plupart des associations détectées significatives sur les caractères foliaires (LAI, indice foliaire) colocalisent à la fois avec des QTL liés à des caractères foliaires et phénologiques (idem pour les associations avec les caractères phénologiques). Ceci illustre que la densité de marquage utilisé en génétique d'association permet possiblement de séparer les locus associés aux caractères foliaires de ceux associés à la phénologie, que l'approche « analyse de liaison » pourrait à tort considérer comme un seul et même locus à effet pleiotropique. Quant aux associations liées à la productivité, les QTL qui colocalisent sont soit liés à la productivité et l'huile pour les chromosomes 01 et 10, ce qui confirme leur intérêt ou soit liés à la phénologie et les caractères foliaires pour le LG09, zone dont l'influence de la tardiveté sur le rendement avait été décrite dans la partie 3. Il est donc probable que l'effet bénéfique sur le rendement de ces allèles ne soit dû qu'à une plus grande tardiveté.

A contrario, pour 72 % des QTL identifiés, des SNP ont été identifiés par génétique d'association dans les mêmes zones. Les colocalisations observées confirment le rôle probable de plusieurs zones. Ainsi, malgré que le LG01 soit caractérisé par la présence de QTL à très larges intervalles, l'un d'entre d'eux, expliquant 20 % de RDTH à GA10\_NI, colocalise avec 2 SNP associés chacun à RDTH dans 2 environnements différents (AI08\_I et VE10) et localisés au niveau de la position la plus probable du QTL.

L'intérêt du LG05 est à nouveau confirmé grâce à la présence de SNP (non polymorphe entre les parents de la population) associés à RDTH (SE10) au niveau d'un cluster de QTL comprenant des QTL de RDT, RDTH, LAD notamment avec des intervalles très courts.

Le rôle du LG10 dans la productivité est consolidé par la présence de QTL et d'associations avec le rendement et ses composantes. Dans cette zone, il avait été observé un nombre

Environnement	Caractère	Chromosome	Inf <sup>1</sup>	Sup <sup>2</sup>	Position max <sup>3</sup>	R <sup>2</sup>	MCQTL test	Effet	Colocalisation Associations				
									Phénologie	Foliaire	Productivité	Huile	Précocité
GA10_NI	LAI_F20	LG01	0.13	14.39	8.2	0.17	4.41	0.04					
GA10_I	PMG	LG01	8.20	60.32	11.7	0.16	4.17	0.27	x	x	x		x
GA10_NI	Huile	LG01	21.41	62.50	55	0.17	4.52	0.14	x	x			x
GA10_NI	RDTH	LG01	28.27	61.25	39.8	0.20	5.34	0.23	x	x	x		
AUZ10_I	F1	LG01	43.06	62.50	58.5	0.21	4.65	0.22		x			
AUZ10_I	M0	LG01	47.06	62.50	54.5	0.18	4.11	0.36		x			
GA10_I	Nbgrains	LG02	28.91	48.98	31.8	0.17	4.27	1541.83	x	x			x
GA10_NI	Nbgrains	LG02	29.15	52.67	34	0.16	4.17	1099.96	x	x			x
GA10_NI	IFF1	LG04	59.90	77.90	63.2	0.17	4.48	1.24	x				
GA10_NI	M3	LG05	37.90	52.57	43.4	0.15	4.07	-0.36	x		x		x
GA10_NI	RDTH	LG05	39.50	46.56	41.5	0.15	4.00	-0.19	x		x		x
GA10_NI	LAI_F20	LG05	39.74	44.19	41.5	0.24	6.49	-0.04	x		x		x
GA10_NI	RDT	LG05	40.02	44.30	41.5	0.21	5.57	-0.45	x		x		x
GA10_NI	LAD	LG05	40.28	41.50	41.3	0.27	7.43	-1.09	x		x		x
GA10_I	F1	LG06	15.75	34.17	20.6	0.21	5.41	-0.27					
GA10_I	Nbgrains	LG06	15.79	26.60	20.6	0.23	5.64	-1811.93					
GA10_NI	M0	LG06	18.72	49.64	29.3	0.15	4.06	-0.20					
AI10_I	F1	LG06	20.28	41.21	26.7	0.15	4.81	-0.13					
AI10_NI	F1	LG06	20.75	30.60	25.9	0.16	5.13	-0.17					
AI10_I	Hauteur	LG06	20.84	41.77	30.6	0.16	5.07	-1.07					
GA10_NI	Hauteur	LG06	24.22	31.55	28.1	0.23	5.91	-1.37					
GA10_I	M0	LG06	24.85	40.06	29.9	0.22	5.85	-0.38					
GA10_NI	F1	LG06	26.33	32.14	29.7	0.22	5.85	-0.23					
AUZ10_I	LAI_F20	LG08	0.00	20.01	0	0.17	4.03	-0.03					x
AUZ10_NI	LAI_F20	LG08	0.00	19.91	8.7	0.19	4.32	-0.02					
AUZ10_NI	LAD	LG08	4.98	19.08	12.6	0.18	4.11	-0.93					
AUZ10_I	F1	LG08	6.64	19.72	9.9	0.20	4.59	-0.20					
AI10_I	Hauteur	LG09	21.38	32.14	30.1	0.17	5.33	1.11	x	x			x
AI10_NI	F1	LG09	28.88	33.26	30.1	0.22	7.03	0.21	x	x		x	x
AI10_I	F1	LG09	29.53	32.55	30.1	0.20	6.23	0.16	x	x		x	x
GA10_I	Hauteur	LG09	30.80	38.97	34.2	0.23	6.10	1.14	x	x		x	x
GA10_NI	Hauteur	LG09	31.90	83.39	34.2	0.17	4.45	1.21	x	x			x
AUZ10_NI	LAI_F40	LG09	46.48	68.77	50.3	0.18	4.30	-0.02					x
GA10_I	Nbgrains	LG09	68.65	75.26	70.5	0.25	6.20	1976.93					
GA10_I	M0M3	LG10	20.38	32.27	25.4	0.18	4.91	0.31	x	x			
GA10_I	RDT	LG10	27.12	39.42	30.1	0.21	5.60	0.82	x				
GA10_NI	RDTH	LG10	27.18	39.53	34.5	0.25	6.62	0.29	x				
GA10_NI	IFF1	LG10	29.33	43.13	41.4	0.17	4.63	0.03	x		x		
GA10_I	Hauteur	LG10	29.49	41.47	35	0.17	4.47	1.02	x				
GA10_I	Nbgrains	LG10	33.60	44.20	38.7	0.19	4.86	1890.33	x	x	x		
GA10_NI	Hauteur	LG10	33.74	46.37	38.9	0.22	5.75	1.60		x	x		
GA10_NI	RDT	LG10	34.22	39.10	35.8	0.33	8.99	0.69					
GA10_NI	PMG	LG10	43.96	48.60	47.5	0.22	6.03	0.90	x	x	x		
GA10_I	Nbgrains	LG10	54.49	55.23	55.2	0.33	8.36	-6876.89	x	x	x		
GA10_I	Nbgrains	LG12	61.76	65.87	63.7	0.26	6.58	-2013.01					
AUZ10_I	F1M0	LG13	0.00	47.14	20.9	0.18	4.12	0.17		x	x	x	x
GA10_I	Nbgrains	LG13	2.47	7.15	5.6	0.26	6.60	-2041.91			x		
AI10_I	Huile	LG13	56.49	56.85	56.8	0.20	6.35	0.43	x	x	x		
GA10_I	LAD	LG14	29.48	56.51	40.5	0.15	4.11	1.52	x	x	x	x	x
GA10_I	M0	LG14	32.25	46.09	41.3	0.19	5.04	0.34	x	x	x		x
GA10_I	F1	LG14	32.37	42.20	34.2	0.19	4.95	0.25		x	x		x
GA10_I	Nbgrains	LG14	38.07	52.82	48.7	0.16	3.96	1539.08	x	x			
GA10_NI	Hauteur	LG14	46.11	54.61	48	0.27	7.03	1.56					
GA10_I	LAI_F20	LG15	27.45	61.70	47.8	0.15	3.92	-0.04	x	x			x
GA10_NI	F1	LG15	28.30	59.71	43.8	0.16	4.31	-0.19	x	x			x
GA10_I	LAI_F1	LG15	43.61	62.25	47.8	0.16	4.35	-0.06	x	x			x
AUZ10_NI	F1M0	LG15	69.84	76.50	72.8	0.22	5.20	-0.25		x			
AUZ10_NI	PMG	LG16	20.08	89.69	40.3	0.18	4.11	-0.22	x			x	x
AUZ10_I	PS	LG16	21.60	46.03	39.5	0.17	3.92	-0.21				x	

**Table V4 : Statistiques pour les QTL détectés dans la population bi parentale.** « Inf » représente la position inférieure de l'intervalle de confiance du QTL et « Sup » la position supérieure, « Position max » est la position la plus probable et R2, la variance expliquée par le QTL et l'Effet est celui apporté par le parent XRQ.

important de SNP associés avec le modèle Kais. Cette colocalisation confirme, qu'en dehors d'une corrélation possible des fréquences alléliques aux groupes de structure, il existe probablement un effet significatif de ces SNP sur le rendement.

Sur le LG13, un QTL de teneur en huile situé sur un intervalle de 0.4 cM colocalise avec des SNP associés à RDTH sur deux environnements différents à partir du modèle Kais. Le LG13 est caractérisé, tout comme le LG10, par de grands blocs de DL dus à la structure et une inflation du nombre d'associations détectées qu'avec le modèle Kais. Cependant, ces informations confirment la possibilité d'une zone « huile » à ce niveau.

Malgré la présence de plusieurs QTL sur le chromosome 6 (phénologie), aucune association n'a été détectée dans cette zone. Parmi les explications possibles, il pourrait s'agir d'un locus à allèle rare ou corrélé à la structure, qui dans ces deux cas n'aurait pu être détecté via la génétique d'association. L'article décrit le même cas pour la floraison sur le LG14, mais il s'avère que des marqueurs associés à d'autres caractères (PMG, hauteur, autres stades phénologiques et RDT) colocalisent avec ce QTL.

Cette étude nous rappelle que, sans une certaine fréquence minimale, le panel d'association ne permet pas d'avoir suffisamment de puissance pour détecter ce type de signal. De plus, l'effet de la structure est également une source importante de réduction de la puissance des tests. L'intérêt de la population biparentale est donc confirmé mais il est clair qu'il n'est pas possible de mener systématiquement ces deux types d'études en parallèle. Ainsi, de nouveaux dispositifs de type « famille », affranchis des problèmes liés à la structuration, et permettant d'augmenter la fréquence de certains allèles, ont été mis en place (cf. discussion).

---

## Discussion

### *Bilan du dispositif mis en place pour la détection des associations marqueurs-phénotype*

La détermination des bases génétiques des caractères complexes est depuis longtemps une préoccupation majeure pour les généticiens cherchant à améliorer les plantes. L'objectif de ce travail de thèse était d'identifier des régions génomiques impliquées dans la productivité du tournesol sous contraintes hydriques. Pour mener à bien cet objectif, un panel de lignées cultivées et élites dont certaines comportant des segments génomiques provenant d'accessions sauvages, a été évalué phénotypiquement sur un réseau expérimental composé de 17 environnements. Des associations entre marqueurs et phénotypes ont pu être identifiées dans l'ensemble des environnements et pour la plupart des caractères. Les associations étaient cependant plus nombreuses pour les caractères de phénologie que pour ceux liés à la productivité. L'héritabilité en générale plus faible pour les caractères complexes tels que le rendement en est une cause possible. Pourtant, le caractère de teneur en huile, fortement héritable n'a permis de mettre en évidence que très peu de zones génomiques. De plus, les effets associés aux marqueurs significatifs sont souvent très faibles. L'héritabilité manquante, c'est à dire la proportion de variance génétique ne pouvant être expliquée par les marqueurs est un sujet central (« Missing heritability problem »), notamment dans le domaine de la génétique humaine où de nombreuses associations ont été découvertes mais expliquent un faible pourcentage de la diversité phénotypique (Manolio *et al.* 2009). Chez les plantes, les résultats dépendent des espèces étudiées et de l'architecture génétique des caractères. Ainsi, chez *Arabidopsis*, la génétique d'association a permis d'identifier des loci expliquant jusqu'à 45% de la variabilité de la floraison (Huang *et al.*, 2010). Chez le maïs, les conclusions semblent être assez proches de celles de la génétique humaine même si récemment il semblerait qu'une grande part de l'héritabilité manquante ait été expliquée grâce à certains dispositifs (Wallace *et al.*, 2013). L'utilisation d'une population de type NAM (Nested association mapping) dans laquelle 25 populations biparentales sont reliées par un parent commun (B73) a permis d'identifier plusieurs locus (au nombre d'une trentaine dans chaque étude) expliquant plus de 80% de la variance génétique de caractères de floraison, de maladie ou encore d'architecture foliaire (Buckler *et al.* 2009, Kump *et al.* 2011, Tian, *et al.* 2011). Contrairement à la génétique humaine, les plantes offrent donc davantage de possibilités pour créer des populations expérimentales permettant d'obtenir une meilleure puissance statistique.



Certains des facteurs souvent évoqués pour expliquer l'héritabilité manquante, tels que la présence d'allèles rares, l'épigénétique et/ou l'épistasie ont probablement un rôle à jouer dans les résultats que nous avons observé dans cette étude. Cependant, il y a aussi plusieurs facteurs apparents, liés au dispositif expérimental, qui permettent de mettre en évidence une puissance de détection des locus sous-optimale. Les facteurs les plus marquants sont détaillés dans le paragraphe ci-dessous avec lorsque cela est possible des solutions alternatives.

#### - Evaluation en combinaison hybride

Toutes les lignées du panel ont été évaluées en combinaison hybride croisées avec des testeurs le plus souvent de type « élite ». Comparé à une évaluation *per se*, l'utilisation des hybrides a tendance à diminuer la variance génétique, les allèles dominants provenant du testeur pouvant masquer les effets des allèles de la lignée (Melchinger *et al.*, 1998). C'est ce que nous avons pu constater au sein de la population de RIL pour lesquels les deux types d'évaluation (*per se* et hybrides) avaient été menés. De nombreuses études attestent de la faiblesse des corrélations entre valeurs hybrides et *per se* pour des caractères comme le rendement où la part de variance additive est plus faible comparé à d'autres caractères morphologiques ou de développement (Milhaljevic *et al.* 2005, Peng *et al.* 2013). Les lieux où les RILs ont été évalués *per se* ont tout de même permis de confirmer certaines zones identifiées en valeur hybride tout en identifiant de nouvelles.

L'évaluation *per se* du panel d'association aurait peut-être pu augmenter la variance génétique et ainsi la puissance des tests mais aurait également probablement conduit à une interaction génotype - environnement biotique ou abiotique plus forte, étant donné la présence de matériel sauvage et la sensibilité possible aux maladies. De plus, ce dispositif permet de se placer dans un contexte propice à la sélection dont l'objectif est d'obtenir les meilleures combinaisons hybrides. Cependant, l'utilisation des testeurs évoluant dans le temps, on peut se demander si les locus identifiés en association avec le phénotype seront toujours significatifs en interaction avec les allèles d'un autre testeur. Un dispositif avec davantage de testeurs afin d'avoir une meilleure évaluation de l'aptitude générale à la combinaison serait envisageable ou l'utilisation de testeurs apparentés et non apparentés, de manière similaire au dispositif mis en place pour la population de RIL dans notre étude. En effet, le fait de disposer d'autres combinaisons de type R x R ou B x B permettrait d'élargir la variabilité génétique (cf. discussion de l'article).





## - Structuration du panel

Une difficulté supplémentaire provient de la présence à la fois de lignées mâles (R) et de lignées femelles (B) au sein du panel ce qui a nécessité la mise en place d'un dispositif au champ spécifique avec l'utilisation de testeurs différents pour les lignées B et les lignées R. De plus, deux testeurs différents par lignée ont été choisis selon les lieux, ce qui a ajouté un niveau de complexité supérieur. L'effet « paire de testeurs » utilisée pour le panel s'est révélé explicatif d'une part non négligeable de l'interaction génotype-environnement. La question de la stabilité des QTL entre testeurs est ainsi à nouveau posée.

Le génotypage du panel a permis de confirmer cette structuration en deux groupes, qui a ainsi été prise en compte dans les modèles d'association afin de corriger de la présence de faux positifs dus à cette structure provoquant inévitablement une perte de puissance des tests car une diminution des effectifs efficaces. Finalement, la plupart des faux positifs que l'effet de structure a permis d'éliminer étaient situés sur les chromosomes 10 et 13 pour lesquels des gènes sélectionnés comme celui de la ramification ont provoqué des phénomènes d'hitchiking. Concernant les autres chromosomes, le nombre d'associations détectées avec ou sans structure était très semblable. Par contre, c'est l'apparement fort au sein du panel qui a probablement induit le plus de perte de puissance lors des tests d'association. La diversité du panel est donc un facteur essentiel pour obtenir une puissance statistique suffisante d'autant plus lorsque l'architecture génétique est caractérisée par la présence de nombreux locus à effets faibles.

## - Couverture du génome

Le déséquilibre de liaison s'étend en moyenne sur 0.14 cM dans le panel (avec un seuil de  $R^2$  égal à 0.20), après correction de la structure et de l'apparement. Plus de 12 000 marqueurs seraient nécessaires pour couvrir entièrement le génome, c'est-à-dire avoir une chance de tester tous les marqueurs possiblement en DL avec les gènes causaux. Nous n'avons pas atteint cette densité dans notre étude donc il est probable que certaines associations n'aient pu être détectées pour cette raison. Nous avons tout de même identifié davantage de marqueurs associés dans les gènes candidats préalablement définis. Cette stratégie est donc intéressante à mettre en œuvre lorsque la densité de marquage est sous-optimale. Grâce aux nouvelles technologies de séquençage et de génotypage, cette densité est aujourd'hui très facilement accessible. Une puce axiome de 200 000 SNP a d'ores et déjà été déployée sur ce panel et



permettra d'augmenter la puissance de détection des associations. Cependant, les puces de génotypage ne contiennent en général qu'une fraction de SNP détectés à partir d'un sous-ensemble de lignées, ce qui réduit inévitablement la capacité de détecter les allèles rares et le polymorphisme causal (Brachi *et al.*, 2011).

#### - Interaction génotype-environnement

Ce travail de thèse a pu bénéficier d'un dispositif expérimental conséquent composé de trois années et de huit lieux répartis sur les zones de production du tournesol en France. Au total, le panel a été évalué sur 17 environnements pour un ensemble de caractères liés à la phénologie, au maintien de la capacité photosynthétique et à la productivité. Les associations détectées sont la plupart du temps spécifiques à un environnement. L'interaction génotype-environnement est très significative, même lorsqu'on considère les environnements auxquels ont été alloués la même paire de testeurs. Ce résultat est aussi appuyé par le fait que les corrélations entre environnements deviennent quasiment inexistantes lorsque l'apparement entre les lignées n'est pas considéré. L'importance de l'interaction génotype-environnement n'est pas surprenant et constitue un problème récurrent dans l'expression des caractères complexes. A ce jour, la plupart des études de génétique d'association n'ont considéré que la moyenne des caractères pour rechercher leur déterminisme génétique, mais la variance phénotypique constitue également une cible intéressante permettant potentiellement de découvrir les bases génétiques de la plasticité (Korte *et al.*, 2013). C'est dans cet esprit que nous avons essayé de décrire les lignées du panel à travers la réponse au stress hydrique de leur productivité en valeur hybride telle qu'estimée grâce à un modèle de culture. Les paramètres extraits de ces droites de réponse ont permis de détecter quelques locus associés mais à des seuils peu significatifs. Parmi les explications possibles, le caractère intégratif des index choisis est probablement à l'origine d'un manque de puissance statistique. Il existe également une certaine incertitude sur le nombre de jours de stress extrait du modèle SUNFLO. En effet, de nombreux axes d'améliorations sont envisageables pour augmenter la précision de la modélisation du stress. Tout d'abord, il est important d'acquérir les données environnementales d'entrée du modèle les plus fiables possibles. Cette étape s'est révélé assez complexe, notamment pour la caractérisation de la profondeur du sol ou encore l'origine des données météo. De plus, le modèle ayant la particularité de prendre en compte les réponses génotypiques des variétés témoins dans l'évaluation du stress, nous aurions pu imaginer un phénotypage plus fin de ces variétés.



## ***Bilan des réponses du panel à la sécheresse***

La disponibilité en eau est primordiale pour le rendement notamment aux phases les plus sensibles du cycle. Tels que caractérisés par le modèle SUNFLO, les niveaux de stress que nous avons pu observer étaient très différents entre les environnements : cumulé tout au long du cycle, le nombre de jour de stress varie de 5 à 50 jours. Concernant les scénarios de stress, les environnements sont par contre assez semblables, caractérisés par une absence de stress hydrique avant la floraison, un stress autour de la floraison pour la plupart des environnements non irrigués et un stress post-floraison plus aléatoire. Quant au stress thermique, il a été principalement provoqué par le froid et sur la plupart des environnements en début de cycle. Au final, peu de symptômes de stress sévère, tel que le flétrissement, ont été observés. Quant à la mise en place de conditions irriguées et non irriguées sur le même lieu, elle ne s'est pas révélée pertinente, le différentiel de stress n'étant pas toujours visible. Ces résultats illustrent toute la difficulté liée au contrôle de l'application du stress hydrique. D'autres stratégies ont été mises en place pour mieux contrôler la survenue, l'intensité et la durée du stress, comme par exemple le décalage de semis qui permet de faire coïncider le stress à différents stades clés ou encore l'utilisation des contre-saisons qui permet d'éviter la présence des précipitations (Blum *et al.*, 2011). La caractérisation du statut hydrique des plantes afin de contrôler plus finement l'irrigation est également une stratégie importante, rendue aujourd'hui davantage possible en condition de plein champ grâce au phénotypage haut-débit. Cette caractérisation peut aussi s'appuyer sur la recherche de bio-marqueurs utilisant les données de séquençage de nouvelle génération (Marchand *et al.*, 2013).

Si les scénarios de stress semblent finalement assez proches, peu d'associations marqueur-phénotype sont communes à plusieurs environnements (comme mentionné dans le paragraphe précédent). Par contre certains locus sont associés avec différents caractères selon les lieux, associations pour lesquelles les effets sont corrélés entre les environnements. Ces corrélations permettent de mettre en évidence une des stratégies mise en place par certains génotypes pour construire leur rendement malgré le stress hydrique, l'évitement. En effet, alors que certains allèles ayant un effet favorable pour des caractères de productivité étaient corrélés avec la tardiveté dans des environnements peu stressés, d'autres locus ont montré une corrélation positive de l'effet favorable pour le rendement avec la précocité. Parmi les caractères phénotypés, certains ont démontré un intérêt particulier pour l'étude du rendement sous contrainte hydrique. Ainsi l'indice foliaire mesuré à 40 jours a permis de discriminer



davantage les lignées du panel sur les environnements stressés (meilleure héritabilité qu'en condition non stressée). Bien corrélé au rendement, cet indice foliaire a permis la détection de QTL adaptatifs. Ces résultats correspondent à l'une des stratégies les plus avantageuses pour le rendement, décrite en introduction, qui est le maintien de la capacité photosynthétique et donc de la transpiration. La corrélation des effets alléliques pour de nombreux caractères suggère l'intérêt d'une approche multi-caractères. L'analyse combinée de l'ensemble des caractères dans un seul modèle de génétique d'association pourrait peut-être apporter davantage de puissance statistique comparé à une approche d'analyse caractère par caractère (Korte et al., 2013). De plus, cela permettrait de pouvoir répondre à une question restée en suspens : ces locus sont-ils vraiment pléiotropiques ou s'agit-il d'une série de locus très proches, chacun affectant un caractère différent ?

### ***Utilisation des résultats et applications en sélection***

Cette étude figure parmi les premières analyses d'association sur le tournesol. Nous avons pu, grâce à la recherche du polymorphisme à l'échelle du génome, appréhender l'architecture génétique de plusieurs caractères. Ce premier résultat est d'intérêt car il permet d'orienter les dispositifs de futures études d'association, qui pourront également espérer bénéficier de nouveaux modèles statistiques tels que les modèles mixtes multi locus (Segura *et al.*, 2012).

Certains désavantages liés à la génétique d'association, comme par exemple la structure, ont été contrebalancés grâce à l'utilisation d'une population biparentale. A l'inverse, la génétique d'association a permis d'obtenir une meilleure résolution pour certaines régions.

Au final, une liste de locus d'intérêt appuyée sur les résultats combinés de ces deux approches, a pu être établie et certains gènes candidats suggérés par homologie de séquences avec *Arabidopsis*. Hormis les perspectives à court terme de densification des zones détectées en marqueurs, d'autres validations sont nécessaires avant une application possible en sélection. En effet, la stabilité des allèles favorables détectés pose question étant donné la spécificité des environnements et du fond génétique (avec notamment la contribution des testeurs) et la faiblesse des effets.

Face à ces contraintes liées au déterminisme génétique des caractères complexes, d'autres approches récemment appliquées aux plantes, telles que la sélection génomique (Meuwissen *et al.*, 2001), semblent plus appropriées dans une perspective appliquée. Cependant, la question de l'interaction génotype-environnement n'est pas encore réglée, y compris dans ces modèles.





---

## Conclusion

La compréhension des déterminismes génétiques du rendement sous contrainte hydrique est un sujet ambitieux de par la complexité de ce caractère et les interactions multiples existant entre les génotypes et leur environnement. Dans cette étude, nous avons bénéficié d'un dispositif conséquent pour répondre à cette question. Il nous a permis de mieux appréhender les caractères utiles et surtout de donner une place centrale à la caractérisation des environnements pour expliquer la réponse des génotypes. Comme en témoignent les résultats d'association, la route est encore longue avant une utilisation efficace de ces marqueurs en sélection mais plusieurs voies d'amélioration ont pu être proposées.



---

## Références

- Alcamo, J., Dronin, N., Endejan, M., Golubev, G., Kirilenko, A., (2007). A new assessment of climate change impacts on food production shortfalls and water availability in Russia. *Global Environmental Change* 17 (3–4), 429–444.
- Aranzana, M.J., Kim, S., Zhao, K., Bakker, E., Horton, M., Jakob, K., Lister, C., Molitor, J., Shindo, C., Tang, C., et al. (2005). Genome-Wide Association Mapping in *Arabidopsis* Identifies Previously Known Flowering Time and Pathogen Resistance Genes. *PLoS Genetics* 1, e60.
- Arcade, A., Labourdette, A., Falque, M., Mangin, B., Chardon, F., Charcosset, A., Joets, J. (2004) BioMercator: integrating genetic maps and QTL towards discovery of candidate genes. *Bioinformatics* 20, 2324–2326
- Astle, W., and Balding, D.J. (2009). Population Structure and Cryptic Relatedness in Genetic Association Studies. *Statistical Science* 24, 451–471.
- Atwell, S., Huang, Y.S., Vilhjálmsson, B.J., Willems, G., Horton, M., Li, Y., Meng, D., Platt, A., Tarone, A.M., Hu, T.T., et al. (2010). Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature* 465, 627–631.
- Baack, E.J., Sapir, Y., Chapman, M.A., Burke, J.M., and Rieseberg, L.H. (2007). Selection on domestication traits and quantitative trait loci in crop-wild sunflower hybrids. *Molecular Ecology* 17, 666–677.
- Bachlava, E., Taylor, C.A., Tang, S., Bowers, J.E., Mandel, J.R., Burke, J.M., and Knapp, S.J. (2012). SNP Discovery and Development of a High-Density Genotyping Array for Sunflower. *PLoS ONE* 7, e29814.
- Barker, T.H., Campos, H., Cooper, M., Dolan, D., Edmeades, G.O., Habben, J., Schussler, J., Wright D and Zinselmeier, C. (2005). Improving drought tolerance in maize. *Plant Breeding Reviews* 25, 173–253.
- Beló, A., Zheng, P., Luck, S., Shen, B., Meyer, D.J., Li, B., Tingey, S., and Rafalski, A. (2008). Whole genome scan detects an allelic variant of *fad2* associated with increased oleic acid levels in maize. *Mol. Genet. Genomics* 279, 1–10.
- Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate — a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* 57, 289–300.
- Beavis, W.D. 1998. QTL analysis: Power, precision, and accuracy. pp. 145-161. In A.H. Paterson (ed.) *Molecular dissection of complex traits*. CRC Press, Boca Raton, FL.
- Berry, S.T., Allen, R.J., Barnes, S.R., and Caligari, P.D.S. (1994). Molecular marker analysis of *Helianthus annuus* L. 1. Restriction fragment length polymorphism between inbred lines of cultivated sunflower. *Theoret. Appl. Genetics* 89, 435–441.
- Bert, P.-F., Jouan, I., de Labrouhe, D.T., Serre, F., Philippon, J., Nicolas, P., and Vear, F. (2003). Comparative genetic analysis of quantitative traits in sunflower (*Helianthus annuus* L.). 2. Characterisation of QTL involved in developmental and agronomic traits. *Theoretical and Applied Genetics* 107, 181–189.



- Blackman, B.K., Rasmussen, D.A., Strasburg, J.L., Raduski, A.R., Burke, J.M., Knapp, S.J., Michaels, S.D., and Rieseberg, L.H. (2011). Contributions of Flowering Time Genes to Sunflower Domestication and Improvement. *Genetics* 187, 271–287.
- Blum, A.(1997) Constitutive traits affecting plant performance under drought stress. In: Edmeades GO, Banziger M, Mickelson HR et al (eds) Developing drought and low N tolerant maize.CIMMYT, El Batan
- Boer, M.P., Wright, D., Feng, L., Podlich, D.W., Luo, L., Cooper, M., van Eeuwijk, F.A. (2007) Amixed-model quantitative trait loci (QTL) analysis for multiple-environment trial data using environmental covariables for QTL-by-environment interactions, with an example in maize. *Genetics*, 177, 1801-1813.
- Bolaños, J. and Edmeades, G.O., (1993). Eight cycles of selection for drought tolerance in lowland and tropical maize. 2. Responses in reproductive behavior. *Field Crops Research* 31, 253–268.
- Borrell, A.K., Hammer, G.L. and Douglas, A.C.L.,(2000). Does maintaining green leaf area in sorghum improve yield under drought? I.Leaf growth and senescence. *Crop Science* 40,1026–1037.
- Boudsocq, M., and Sheen, J. (2013). CDPKs in immune and stress signaling. *Trends Plant Sci* 18, 30–40.
- Bouzidi, M.F., Badaoui, S., Cambon, F., Vear, F., De Labrouhe, D.T., Nicolas, P., and Mouzeyar, S. (2002). Molecular analysis of a major locus for resistance to downy mildew in sunflower with specific PCR-based markers. *Theor. Appl. Genet.* 104, 592–600.
- Bowers, J.E., Bachlava, E., Brunick, R.L., Rieseberg, L.H., Knapp, S.J., and Burke, J.M. (2012). Development of a 10,000 Locus Genetic Map of the Sunflower Genome Based on Multiple Crosses. *G3* 2, 721–729.
- Brachi, B., Faure, N., Horton, M., Flahauw, E., Vazquez, A., Nordborg, M., Bergelson, J., Cuguen, J., and Roux, F. (2010). Linkage and Association Mapping of *Arabidopsis thaliana* Flowering Time in Nature. *PLoS Genet* 6, e1000940.
- Breseghele, F., and Sorrells, M.E. (2006). Association Mapping of Kernel Size and Milling Quality in Wheat (*Triticum aestivum* L.) Cultivars. *Quality* 1177, 1165–1177.
- Buckler ES, Holland JB, Bradbury PJ, Acharya C, Brown P, Browne C, Ersoz E, Flint-Garcia S, Garcia A, Glaubitz JC, Goodman MM, Harjes C, Guill K, Kroon DE, Larsson S, Lepak NK, Li H, Mitchell SE, Pressoir G, Peiffer JA, Oropeza Rosas M, Rocheford TR, Romay MC, Romero S, Salvo S, Villeda HS, da Silva HS, Sun Q, Tian F, Upadyayula N, Ware D, Yates H, Yu J, Zhang Z, Kresovich S, McMullen MD. (2009) The genetic architecture of maize flowering time. *Science* 325:714-718.
- Burke, J.M. (2005). Genetic Consequences of Selection During the Evolution of Cultivated Sunflower. *Genetics* 171, 1933–1940.
- Burke, J.M., Lai, Z., Salmaso, M., Nakazato, T., Tang, S., Heesacker, A., Knapp, S.J., and Rieseberg, L.H. (2004). Comparative mapping and rapid karyotypic evolution in the genus *Helianthus*. *Genetics* 167, 449–457.
- Burke, J.M., Tang, S., Knapp, S.J., and Rieseberg, L.H. (2002). Genetic analysis of sunflower domestication. *Genetics* 161, 1257–1267.



- Cabrera-Bosquet, L., Crossa, J., von Zitzewitz, J., Serret, M.D., and Araus, J.L. (2012). High-throughput phenotyping and genomic selection: the frontiers of crop breeding converge. *J Integr Plant Biol* 54, 312–320.
- Cadic, E., Coque, M., Vear, F., Grezes-Besset, B., Pauquet, J., Piquemal, J., Lippi, Y., Blanchard, P., Romestant, M., Pouilly, N., et al. (2013). Combined linkage and association mapping of flowering time in Sunflower (*Helianthus annuus* L.). *Theor. Appl. Genet.* 126, 1337–1356.
- Cairns, J.E., Sanchez, C., Vargas, M., Ordoñez, R., and Araus, J.L. (2012). Dissecting maize productivity: ideotypes associated with grain yield under drought stress and well-watered conditions. *J Integr Plant Biol* 54, 1007–1020.
- Casadebaig, P., Debaeke, P., and Lecoeur, J. (2008). Thresholds for leaf expansion and transpiration response to soil water deficit in a range of sunflower genotypes. *European Journal of Agronomy* 28, 646–654.
- Casadebaig, P., Guilioni, L., Lecoeur, J., Christophe, A., Champolivier, L., and Debaeke, P. (2011). SUNFLO, a model to simulate genotype-specific performance of the sunflower crop in contrasting environments. *Agricultural and Forest Meteorology* 151, 163–178.
- Cattivelli L., Rizza F., Badeck FW., Mazzucotelli E., Mastrangelo AM, Francia E., Mare`C., Tondelli A., Stanca M.(2008). Drought tolerance improvement in crop plants: An integrated view from breeding to genomics. *Field Crop Research* 105,1-14.
- Ceccarelli S, Grando S, Impiglia A (1998) Choice of selection strategy in breeding barley for stress environments. *Euphytica* 10, 307–318
- Cellier, F., Conéjéro, G., Breitler, J.C., and Casse, F. (1998). Molecular and physiological responses to water deficit in drought-tolerant and drought-sensitive lines of sunflower. Accumulation of dehydrin transcripts correlates with tolerance. *Plant Physiol.* 116, 319–328.
- Chapman, S.C., Cooper, M., Hammer, G.L., Butler, D.G. (2000). Genotype by environment interactions affecting grain sorghum. II. Frequencies of different seasonal patterns of drought stress are related to location effects on hybrid yields. *Australian Journal of Agricultural Research* 51, 209–221.
- Chapman, S.C., Hammer, G.L., and Meinke, H. (1993). A Sunflower Simulation Model: I. Model Development. *Agronomy Journal* 85, 725.
- Charcosset, A., Mangin, B., Moreau, L., Combes, L., Jourjon, M.F, et al., (2000) Heterosis in maize investigated using connected RIL populations, pp. 89-98 in *Quantitative genetics and breeding methods: the way ahead*. INRA, Paris.
- Chen, L., and Storey, J.D. (2006). Relaxed Significance Criteria for Linkage Analysis. *Genetics* 173, 2371–2381.
- Chenu, K., Chapman, S.C., Tardieu, F et al (2009) Simulating the yield impacts of organ-level quantitative trait loci associated with drought response in maize: a “gene-to-phenotype” modeling approach. *Genetics* 183 ,1507–1523
- Chenu, K., Cooper, M., Hammer, G.L., Mathews, K.L., Dreccer, M.F., and Chapman, S.C. (2011). Environment characterization as an aid to wheat improvement: interpreting genotype–environment interactions by modelling water-deficit patterns in North-Eastern Australia. *J. Exp. Bot.* 62, 1743–1755.





- Cheres, M.T., Miller, J.F., Crane, J.M., and Knapp, S.J. (2000). Genetic distance as a predictor of heterosis and hybrid performance within and between heterotic groups in sunflower. *Theoretical and Applied Genetics* 100, 889–894.
- Chervet, B., Vear, F., (1990). Etude des relations entre la precocite du tournesol et son rendement, sa teneur en huile, son developpement et sa morphologie. *Agronomie*, 10, 1
- Chimenti, C., Pearson, J., and Hall, A.. (2002). Osmotic adjustment and yield maintenance under drought in sunflower. *Field Crops Research* 75, 235–246.
- Clark L. J., Price A. H., Steele K. A., Whalley W. R. (2008). Evidence from near-isogenic lines that root penetration increases with root diameter and bending stiffness in rice. *Funct. Plant Biol.* 35, 1163–1171
- Cockram, J., White, J., Zuluaga, D.L., Smith, D., Comadran, J., Macaulay, M., Luo, Z., Kearsley, M.J., Werner, P., Harrap, D., et al. (2010). Genome-wide association mapping to candidate polymorphism resolution in the unsequenced barley genome. *Proceedings of the National Academy of Sciences* 107, 21611–21616.
- Collins, N.C., Tardieu, F., Tuberosa, R. (2008) Quantitative trait loci and crop performance under abiotic stress: where do we stand? *Plant Physiol* 147, 469–486
- Condon, A.G., Richards, R.A., Rebetzke, G.J. (2004) Breeding for high water-use efficiency. *J Exp Bot* 55:2447–2460
- Condon, A.G., Richards, R.A., Rebetzke, G.J., Farquhar, G.D., (2002). Improving intrinsic water use efficiency and crop yield. *Crop Sci.* 42, 122–131.
- Connor, D.J., Jones, T.R. and Palta, J.A. (1985) Response of sunflower to strategies of irrigation. I. Growth, yield and the efficiency of water-use. *Field Crops Research* 10, 15-26
- Coque, M., S. Mesmildrey, M. Romestant, B. Grezes-Besset, F. Vear, N.B. Langlade, and P. Vincourt. 2008. Sunflower nested core collections for association studies and phenomics. p. 725-728. In: Proc. 17th Int. Sunfl. Conf., Córdoba, Spain. Int. Sunfl. Assoc., Paris, France.
- Corder, E.H., Saunders, A.M., Strittmatter, W.J, Schmechel, D.E., et al. (1993) Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer’s disease in late onset families. *Science.* 261, 921–923
- Courtois, B., McLaren, G., Sinha, P.K., Prasad, K., Yadava, R., et al. (2000) Mapping QTLs associated with drought avoidance in upland rice. *Mol Breed* 6, 55–66.
- Crouzillat, D., Canal, L., Perrault, A., Ledoigt, G., Vear, F. and Serieys, (1991), Cytoplasmic male sterility in sunflower: Comparison of molecular biology and genetic studies. *Pl. Mol. Biol.*, 16, 415-426.
- D. Butler, B. R. Cullis, A. R. Gilmour and B. J. Gogel (2007). Analysis of Mixed Models for S Language Environments: ASReml-R Reference Manual. (Brisbane: DPI&F Publications)
- Darvishzadeh, R., Pirzad, A., Hatami-Maleki, H., Poormohammad-Kiani, S., and Sarrafi, A. (2010). Evaluation of the reaction of sunflower inbred lines and their F1 hybrids to drought conditions using various stress tolerance indices. *Spanish Journal of Agricultural Research* 8, 1037–1046.
- De la Vega, A.J., Cantore, M.A., Sposaro, M.M., Trápani, N., López Pereira, M., and Hall, A.J. (2011). Canopy stay-green and yield in non-stressed sunflower. *Field Crops Research* 121, 175–185.



- De la Vega, A.J., and Hall, A.J. (2002). Effects of Planting Date, Genotype, and Their Interactions on Sunflower Yield. *Crop Science* 42, 1202.
- Denis, J.B., et Vincourt P.(1982) Panorama des méthodes statistiques d'analyse des interactions génotype x milieu. *Agronomie*, 2, 3, 219- 230
- Devlin, B., and Roeder, K. (1999). Genomic control for association studies. *Biometrics* 55, 997–1004.
- Dezar, C.A., Gago, G.M., Gonzalez, D.H. and Chan, R.L. (2005a) Hahb-4, a sunflower homeobox-leucine zipper gene, is a developmental regulator and confers drought tolerance to *Arabidopsis thaliana* plants. *Transgenic Res.* 14, 429–440.
- Dezar, C.A., Fedrigo, G.V. and Chan, R.L. (2005b) The promoter of the sunflower HD-Zip protein gene Hahb-4 directs tissue-specific expression
- Dong, J., and Bergmann, D.C. (2010). Stomatal patterning and development. *Curr. Top. Dev. Biol.* 91, 267–297.
- Duvick, D.N. (1997). Heterosis : feeding people and protecting natural resources. CYMMIT Symposium.
- Eathington, S.R., Crosbie, T.M., Edwards, M.D., Reiter, R.S., and Bull, J.K. (2007). Molecular Markers in a Commercial Breeding Program. *Crop Science* 47, S–154.
- Ebrahimi, A., Maury, P., Berger, M., Kiani, S.P., Nabipour, A., Shariati, F., Grieu, P., and Sarrafi, A. (2008). QTL mapping of seed-quality traits in sunflower recombinant inbred lines under different water regimes. *Genome* 51, 599–615.
- Ersoz, E.S., Yu, J., and Buckler, E.S. (2007). Applications of Linkage Disequilibrium and Association Mapping in Crop Plants. In *Genomics-Assisted Crop Improvement*, R.K. Varshney, and R. Tuberosa, eds. (Springer Netherlands), pp. 97–119.
- Evanno, G., Regnaut, S., and Goudet, J. (2005). Detecting the number of clusters of individuals using the software structure: a simulation study. *Molecular Ecology* 14, 2611–2620.
- Falush, D., Stephens, M., and Pritchard, J.K. (2003). Inference of Population Structure Using Multilocus Genotype Data: Linked Loci and Correlated Allele Frequencies. *In Practice* 1587, 1567–1587.
- Famoso, A.N., Zhao, K., Clark, R.T., Tung, C., Wright, M.H., Kochian, L.V., and Mccouch, S.R. (2011). Genetic Architecture of Aluminum Tolerance in Rice ( *Oryza sativa* ) Determined through Genome-Wide Association Analysis and QTL Mapping. *PLoS Genetics* 7.
- Frahm, M.A., Rosas, J.C., Mayek-Perez, N., et al (2004) Breeding beans for resistance to terminal drought in the Lowland tropics. *Euphytica* 136, 223–232
- Federer, W. T. (1956). Augmented (or hoonuiaku) designs. *Hawaiian Planters' Record* LV(2):I9I-208
- Finlay, K.W., and Wilkinson, G.N. (1963). The analysis of adaptation in a plant-breeding programme. *Australian Journal of Agricultural Research* 14, 742–754.
- Flint-Garcia, S.A., Thornsberry, J.M., S, E., and Iv, B. (2003). Structure of linkage disequilibrium in plants. *Annual Review of Plant Biology* 54, 357–374.



- Fusari, C.M., Lia, V.V., Hopp, H.E., Heinz, R.A., and Paniego, N.B. (2008). Identification of Single Nucleotide Polymorphisms and analysis of Linkage Disequilibrium in sunflower elite inbred lines using the candidate gene approach. *BMC Plant Biology* 8, 7.
- Gallais, A., Bannerot, H. (1992). Amélioration des espèces végétales cultivées. Objectifs et critères de sélection. INRA. Paris. 687 p.
- Gao, X., Becker, L.C., Becker, D.M., Starmer, J.D., and Province, M.A. (2010). Avoiding the high Bonferroni penalty in genome-wide association studies. *Genet Epidemiol* 34, 100–105.
- Gauch, H.G. 1988. Model selection and validation for yield trials with interaction. *Biometrics* 44, 705–715.
- Gentzbittel, L., Zhang, Y.-X., Vear, F., Griveau, B., and Nicolas, P. (1994). RFLP studies of genetic relationships among inbred lines of the cultivated sunflower, *Helianthus annuus* L.: evidence for distinct restorer and maintainer germplasm pools. *Theoretical and Applied Genetics* 89, 419–425.
- Gilmour, A.R., B.R. Cullis, and A.P. Verbyla. (1997). Accounting for natural and extraneous variation in the analysis of field experiments. *Journal of Agricultural, Biological, and Environmental Statistics* 2, 269-293.
- Gimenez, C. and E. Fereres. (1986). Genetic variability in sunflower cultivars under drought. II. Growth and water relations. *Australian Journal of Agricultural Research* 37, 583-97.
- Giordani, T., Buti, M., Natali, L., Pugliesi, C., Cattonaro, F., Morgante, M., and Cavallini, A. (2011). An analysis of sequence variability in eight genes putatively involved in drought response in sunflower (*Helianthus annuus* L.). *Theor. Appl. Genet.* 122, 1039–1049.
- Hamblin, M.T., Warburton, M.L., and Buckler, E.S. (2007). Empirical Comparison of Simple Sequence Repeats and Single Nucleotide Polymorphisms in Assessment of Maize Diversity and Relatedness. *Structure*.
- Hammer, G.L. et al. (2005) Trait physiology and crop modelling as a framework to link phenotypic complexity to underlying genetic systems. *Aust. J. Agric. Res.* 56, 947–960
- Harjes, C.E., Rocheford, T.R., Bai, L., Brutnell, T.P., Kandianis, C.B., Sowinski, S.G., Stapleton, A.E., Vallabhaneni, R., Williams, M., Wurtzel, E.T., et al. (2008). Natural Genetic Variation in Lycopene Epsilon Cyclase Tapped for Maize Biofortification. *Science* 319, 330–333.
- Harris, K., Subudhi, P., Borrell, A., Jordan, D., Rosenow, D., Nguyen, H., Klein, P., Klein, R., and Mullet, J. (2006). Sorghum stay-green QTL individually reduce post-flowering drought-induced leaf senescence. *Journal of Experimental Botany* 58, 327–338.
- Hayes, B.J., Bowman, P.J., Chamberlain, A.J., and Goddard, M.E. (2009). Invited review: Genomic selection in dairy cattle: progress and challenges. *J. Dairy Sci.* 92, 433–443.
- Heiser, C.B., Smith, D.M., Clevenger, S.B. and Martin W.C. (1969). The North American Sunflower (*Helianthus*). *Memories Torrey Botanical Club* 22, 1-218.
- Hernandez-Segundo E, Capettini F, Trethowan R, van Ginkel M, Mejia A, Carballo A, Crossa J, Vargas M, Balbuena- Melgarejo A. (2009). Mega-environment identification for barley based on twenty-seven years of global grain yield data. *Crop Science* 49, 1705–1718.
- Hodson, D.P. and White, J.W. (2007). Use of spatial analyses for global characterization of wheat-based production systems. *Journal of Agricultural Science* 145, 115–125.



- Hongtrakul, V., Huestis, G.M., and Knapp, S.J. (1997). Amplified fragment length polymorphisms as a tool for DNA fingerprinting sunflower germplasm: genetic diversity among oilseed inbred lines. *Theoretical and Applied Genetics* 95, 400–407.
- Hsieh, T.H., Lee, J.T., Yang, P.T., Chiu, L.H., Charng, Y., Wang, Y.C., and Chan, M.-T. (2002). Heterology expression of the Arabidopsis C-repeat/dehydration response element binding factor 1 gene confers elevated tolerance to chilling and oxidative stresses in transgenic tomato. *Plant Physiol.* 129, 1086–1094.
- Hu., H, Dai., M, Yao, J., Xiao, B., and Li, X. (2006) Overexpressing a NAM, ATAF, and CUC (NAC) transcription factor enhances drought resistance and salt tolerance in rice. *PNAS* Vol. 103, 12987-12992.
- Huang ,W.L., Lee, C.H., Chen, Y.R. (2012) Levels of endogenous abscisic acid and indole-3-acetic acid influence shoot organogenesis in callus cultures of rice subjected to osmotic stress. *Plant Cell Tiss Organ Cult* 108,257–263
- Huang, X., Zhao, Y., Wei, X., Li, C., Wang, A., Zhao, Q., Li, W., Guo, Y., Deng, L., Zhu, C., et al. (2011). Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. *Nature Genetics* 44, 32–39.
- Hill., W.G., Weir, B.S. (1988) Variances and covariances of squared linkage disequilibria in finite populations. *Theor Popul Biol* 33, 54–78
- Ingvarsson, P.K. (2005). Nucleotide Polymorphism and Linkage Disequilibrium Within and Among Natural Populations of European Aspen (*Populus tremula* L., Salicaceae). *Genetics* 169, 945–953.
- Ingvarsson, P.K., and Street, N.R. (2011). Association genetics of complex traits in plants. *New Phytologist* 189, 909–922.
- Ito, Y., Katsura, K., Maruyama, K., Taji, T., Kobayashi, M., Seki, M., Shinozaki, K., and Yamaguchi-Shinozaki, K. (2006). Functional analysis of rice DREB1/CBF-type transcription factors involved in cold-responsive gene expression in transgenic rice. *Plant Cell Physiol.* 47, 141–153.
- Jarvis, A., Ramírez, J., Anderson, B., Leibing, C., Aggarwal, P. (2010). Scenarios of climate change within the context of agriculture. In: Reynolds, M.P. (ed.). *Climate change and crop production* CABI climate change. CABI, Wallingford, GB. v. 1, p. 1-38.
- Jestin, C., Lodé, M., Vallée, P., Domin, C., Falentin, C., Horvais, R., Coedel, S., Manzanares-Dauleux, M.J., and Delourme, R. (2011). Association mapping of quantitative resistance for *Leptosphaeria maculans* in oilseed rape (*Brassica napus* L.). *Mol Breeding* 27, 271–287.
- Johnstone, I., M., Perry, P., O., Ma , Z., and Shahram , M.(2009). RMT: Distributions, Statistics and Tests derived from Random Matrix Theory. R package version 0.2.
- Johnson, C.D., Chary, S.N., Chernoff, E.A., Zeng, Q., Running, M.P., and Crowell, D.N. (2005). Protein Geranylgeranyltransferase I Is Involved in Specific Aspects of Abscisic Acid and Auxin Signaling in Arabidopsis. *Plant Physiol.* 139, 722–733.
- Jones ES, Sullivan H, Bhatramakki D, Smith JSC (2007) A comparison of simple sequence repeat and single nucleotide polymorphism marker technologies for the genotypic analysis of maize (*Zea mays*L.). *Theor Appl Genet* 115,361–371





- Jouffret, P., Labalette, F., Thibierge, J. (2011). Atouts et besoins en innovations du tournesol pour une agriculture durable. *Innovations Agronomiques*, 14,1-17.
- Jung, M., Ching, A., Bhatramakki, D., Dolan, M., Tingey, S., Morgante, M., and Rafalski, A. (2004). Linkage disequilibrium and sequence diversity in a 500-kbp region around the *adh1* locus in elite maize germplasm. *Theor. Appl. Genet.* 109, 681–689.
- Kamran, A., and Asif, M. (2011). Climate Change and Crop Production. *Crop Science* 51, 2299.
- Kane, N.C., Burke, J.M., Marek, L., Seiler, G., Vear, F., Baute, G., Knapp, S.J., Vincourt, P., and Rieseberg, L.H. (2013). Sunflower genetic, genomic and ecological resources. *Molecular Ecology Resources* 13, 10–20.
- Kane, N.C., Gill, N., King, M.G., Bowers, J.E., Berges, H., Gouzy, J., Bachlava, E., Langlade, N.B., Lai, Z., Stewart, M., et al. (2011). Progress towards a reference genome for sunflower. *Botany* 89, 429–437.
- Kane, N.C., and Rieseberg, L.H. (2007). Selective Sweeps Reveal Candidate Genes for Adaptation to Drought and Salt Tolerance in Common Sunflower, *Helianthus annuus*. *Genetics* 175, 1823–1834.
- Kang, H.M., Zaitlen, N.A., Wade, C.M., Kirby, A., Heckerman, D., Daly, M.J., and Eskin, E. (2008). Efficient Control of Population Structure in Model Organism Association Mapping. *Genetics* 178, 1709–1723.
- Kang, H.M., Sul, J.H., Service, S.K., Zaitlen, N.A., Kong, S., Freimer, N.B., Sabatti, C., and Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nat Genet* 42, 348–354.
- Kantar, M., Unver, T., and Budak, H. (2010). Regulation of barley miRNAs upon dehydration stress correlated with target gene expression. *Funct. Integr. Genomics* 10, 493–507.
- Kiani, S.P., Talia, P., Maury, P., Grieu, P., Heinz, R., Perrault, A., Nishinakamasu, V., Hopp, E., Gentzittel, L., Paniego, N., et al. (2007). Genetic analysis of plant water status and osmotic adjustment in recombinant inbred lines of sunflower under two water treatments. *Plant Science* 172, 773–787.
- Kiers, H.A.L.(1997).Weighted least squares fitting using ordinary least squares algorithms.*Psychometrika* 62, 251–266.
- Kijne, J. W., Barker, Randolph, Molden, D. J. (Eds), 2003. *Water Productivity in Agriculture: Limits and Opportunities for Improvement*, CABI, UK, 332 pp.
- Kinman, M.L., (1970) Letter to participants. *Proceedings of the 4th International Sunflower Conference, Memphis, USA*, 4,181–183.
- Kohli, A., Sreenivasulu, N., Lakshmanan, P., and Kumar, P.P. (2013). The phytohormone crosstalk paradigm takes center stage in understanding how plants respond to abiotic stresses. *Plant Cell Rep* 32, 945–957.
- Kolkman, J.M., Berry, S.T., Leon, A.J., Slabaugh, M.B., Tang, S., Gao, W., Shintani, D.K., Burke, J.M., and Knapp, S.J. (2007). Single Nucleotide Polymorphisms and Linkage Disequilibrium in Sunflower. *Genetics* 177, 457–468.
- Korte, A., Vilhjálmsson, B.J., Segura, V., Platt, A., Long, Q., and Nordborg, M. (2012). A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nat. Genet.* 44, 1066–1071.



- Kramer PJ and Boyer JS (1995). Water relations of plants and soils. Academic Press, San Diego, California, USA, 512 pp.
- Kreps, J.A., Wu, Y., Chang, H.-S., Zhu, T., Wang, X., and Harper, J.F. (2002). Transcriptome changes for Arabidopsis in response to salt, osmotic, and cold stress. *Plant Physiol.* 130, 2129–2141.
- Kulwal, P., Ishikawa, G., Benscher, D., Feng, Z., Yu, L.-X., Jadhav, A., Mehetre, S., and Sorrells, M.E. (2012). Association mapping for pre-harvest sprouting resistance in white winter wheat. *Theoretical and Applied Genetics* 125, 793–805.
- Kumar, A.P., Boualem, A., Bhattacharya, A., Parikh, S., Desai, N., Zambelli, A., Leon, A., Chatterjee, M., and Bendahmane, A. (2013). SMART–Sunflower Mutant population And Reverse genetic Tool for crop improvement. *BMC Plant Biology* 13, 38.
- Kump, K.L., Bradbury, P.J., Wissler, R.J., Buckler, E.S., Belcher, A.R., Oropeza-Rosas, M.A., Zwonitzer, J.C., Kresovich, S., McMullen, M.D., Ware, D., et al. (2011). Genome-wide association study of quantitative resistance to southern leaf blight in the maize nested association mapping population. *Nat Genet* 43, 163–168.
- Lai, Z., Zou, Y., Kane, N.C., Choi, J.H., Wang, X., Rieseberg, L.H. (2012). Preparation Of Normalized cDNA Libraries For 454 Titanium Transcriptome Sequencing. *Methods In Molecular Biology* (clifton, Nj) 888:119
- Lander, E.S., and B, D.B. (1989). Mapping Mendelian Factors Underlying Quantitative Traits Using RFLP Linkage Maps. *Genetics*.
- Leclercq, P. (1969). Une stérilité cytoplasmique chez le tournesol. *Ann Amélior Plant* 19,99–106
- Levitt, J. (1980) Response of plants to environmental stresses water, radiation salt and other stresses. Academic, New York
- Li, Y., Huang, Y., Bergelson, J., Nordborg, M., and Borevitz, J.O. (2010). Association mapping of local climate-sensitive quantitative trait loci in *Arabidopsis thaliana*. *PNAS* 201007431.
- Liu, A. (2006). Patterns of Nucleotide Diversity in Wild and Cultivated Sunflower. *Genetics* 173, 321–330.
- Liu, A., Burke, J.M. (2006). Patterns of nucleotide diversity in wild and cultivated sunflower. *Genetics* 173, 321–330
- Lobell, D.B., Schlenker, W., and Costa-Roberts, J. (2011). Climate trends and global crop production since 1980. *Science* 333, 616–620.
- Ludlow, M.M. and Muchow, R.C. (1990) A critical evaluation of traits for improving crop yields in water-limited environments. *Adv.Agron.* 43,107-153.
- Maccaferri, M., Sanguineti, M.C., Demontis, A., El-Ahmed, A., Garcia del Moral, L., Maalouf, F., Nachit, M., Nserallah, N., Ouabbou, H., Rhouma, S., et al. (2011). Association mapping in durum wheat grown across a broad range of water regimes. *J. Exp. Bot.* 62, 409–438.
- Mackay, I., and Powell, W. (2007). Methods for linkage disequilibrium mapping in crops. *Trends in Plant Science* 12, 57–63.



- Maenhout, S., Baets, B., and Haesaert, G. (2009). Marker-based estimation of the coefficient of coancestry in hybrid breeding programmes. *Theoretical and Applied Genetics* 118, 1181–1192.
- Manavella, P.A., Arce, A.L., Dezar, C.A., Bitton, F., Renou, J.P., Crespi, M. and Chan, R.L. (2006) Cross-talk between ethylene and drought signalling pathways is mediated by the sunflower Hahb- 4 transcription factor. *Plant J.* 48, 125–137.
- Mandel, J.R., Dechaine, J.M., Marek, L.F., and Burke, J.M. (2011). Genetic diversity and population structure in cultivated sunflower and a comparison to its wild progenitor, *Helianthus annuus* L. *Theoretical and Applied Genetics* 123, 693–704.
- Mangin, B., Siberchicot, A., Nicolas, S., Doligez, A., This, P., and Cierco-Ayrolles, C. (2012). Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. *Heredity (Edinb)* 108, 285–291.
- Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. *Nature* 461, 747–753.
- Marchand, G., Mayjonade, B., Varès, D., Blanchet, N., Boniface, M.-C., Maury, P., Nambinina, F.A., Burger, P., Debaeke, P., Casadebaig, P., et al. (2013). A biomarker based on gene expression indicates plant water status in controlled and natural environments. *Plant Cell Environ.* 36, 2175–2189.
- Matsui, A., Ishida, J., Morosawa, T., Mochizuki, Y., Kaminuma, E., Endo, T.A., Okamoto, M., Nambara, E., Nakajima, M., Kawashima, M., et al. (2008). Arabidopsis transcriptome analysis under drought, cold, high-salinity and ABA treatment conditions using a tiling array. *Plant Cell Physiol.* 49, 1135–1149.
- Maynard Smith, J., and J. HAICH (1974) The hitchhiking effect of a favourable gene. *Genet. Res.* 23, 23-35.
- Melchinger, A.E., Utz, H.F., and Schön, C.C. (1998). Quantitative trait locus (QTL) mapping using different testers and independent population samples in maize reveals low power of QTL detection and large bias in estimates of QTL effects. *Genetics* 149, 383–403.
- Messmer, R., Fracheboud, Y., Bänziger, M., Vargas, M., Stamp, P., Ribaut, J.-M., 2009. Drought stress and tropical maize: QTL-by-environment interactions and stability of QTLs across environments for yield components and secondary traits. *Theor. Appl. Genet.* 119, 913–930.
- Meuwissen, T.H., Hayes, B.J., and Goddard, M.E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157, 1819–1829.
- Micic, Z., Hahn, V., Bauer, E., Melchinger, A.E., Knapp, S.J., Tang, S., Schon, C.C., (2005). Identification and validation of QTL for *Sclerotinia* midstalk rot resistance in sunflower by selective genotyping. *Theor Appl Genet*, 111, 233–242.
- Mihaljevic, R., C.C. Schön, H.F. Utz, and A.E. Melchinger. (2005). Correlations and QTL correspondence between line per se and testcross performance for agronomic traits in four populations of European maize. *Crop Sci.* 45, 114-122
- Mitchell J.H., Siamhan D., Wamala M.H., Risimeri J.B., Chinyamakobvu E., Henderson S.A. and Fukai S (1998). The use of seedling leaf death score for evaluating of drought resistance of rice. *Field Crops Research* 55, 129–139.



- Moinuddin, Khannu-Chopra R.(2004). Osmotic adjustment in chickpea in relation to seed yield and yield parameters.CropScience 44,449–455
- Mokrani, L., Gentzbittel, L., Azanza, F., Fitamant, L., Al-Chaarani, G., and Sarrafi, A. (2002). Mapping and analysis of quantitative trait loci for grain oil content and agronomic traits using AFLP and SSR in sunflower (*Helianthus annuus* L.). *Theoretical and Applied Genetics* 106, 149–156.
- Monteith JL. (1977). Climate and the efficiency of crop production in Britain. *Philosophical Transactions of the Royal Society of London B*281, 277-297.
- Montes, J., Melchinger, A., and Reif J. (2007). Novel throughput phenotyping platforms in plant genetic studies. *Trends in Plant Science* 12, 433–436.
- Morgan, J.M. (1991) A gene controlling differences in osmoregulation in wheat. *Aust J Plant Physiol* 18, 249–257
- Moriondo, M., Giannakopoulos, C., and Bindi, M. (2010). Climate change impact assessment: the role of climate extremes in crop yield simulation. *Climatic Change* 104, 679–701.
- Morozov, V.K. (1947). Sunflower selection in USSR. Pishchepromizdat, Moscow, pp 1–272
- Muchow, R.C., Cooper, M., Hammer, G.L. (1996). Characterizing environmental challenges using models. In: Cooper M, Hammer GL, eds. *Plant adaptation and crop improvement*. CAB International, Wallingford, 349–364.
- Müller, B., Stich, B., and Piepho, H. (2010) A general method for controlling the genomewidetype I error rate in linkage and association mapping experiments in plants. *Heredity* 106, 825-831.
- Munns, R., Richards, R.A. (2007) Recent advances in breeding wheat for drought and salt stresses. In: Jenks MA, Hasegawa PM, Jain SM (eds) *Advances in molecular breeding toward drought and salt tolerant crops*. Springer, Dordrecht
- Myles, S., Peiffer, J., Brown, P.J., Ersoz, E.S., Zhang, Z., and Costich, D.E. (2009). Association Mapping: Critical Considerations Shift from Genotyping to Experimental Design. *Society* 21, 2194–2202.
- Nagano, M., Ihara-Ohori, Y., Imai, H., Inada, N., Fujimoto, M., Tsutsumi, N., Uchimiya, H., and Kawai-Yamada, M. (2009). Functional association of cell death suppressor, *Arabidopsis* Bax inhibitor-1, with fatty acid 2-hydroxylation through cytochrome b<sub>5</sub>. *Plant J.* 58, 122–134.
- Nemri, A., Atwell, S., Tarone, A.M., Huang, Y.S., Zhao, K., Studholme, D.J., Nordborg, M., and Jones, J.D.G. (2010). Genome-wide survey of *Arabidopsis* natural variation in downy mildew resistance using combined association and linkage mapping. *Proceedings of the National Academy of Sciences* 107, 10302–10307.
- Niu, Y., Xu, Y., Liu, X.-F., Yang, S.-X., Wei, S.-P., Xie, F.-T., and Zhang, Y.-M. (2013). Association mapping for seed size and shape traits in soybean cultivars. *Mol Breeding* 31, 785–794.
- Nordborg, M. (2000). Linkage disequilibrium, gene trees and selfing: an ancestral recombination graph with partial self-fertilization. *Genetics* 154, 923–929.
- Nordborg, M., and Weigel, D. (2008). Next-generation genetics in plants. *Nature* 456, 720–723.
- Passioura, JB.(1977). Grain yield, harvest index and water use of wheat. *Journal of Australian Institute of Agricultural Science* 43,117–120





- Patterson, N., Price, A.L., and Reich, D. (2006). Population Structure and Eigenanalysis. *PLoS Genetics* 2, e190.
- Pellegrineschi, A., Brito, R.M., McLean, S., Hoisington, D.A. (2004) Effect of 2,4-Dichlorophenoxyacetic Acid and NaCl on the Establishment of Callus and Plant Regeneration in Durum and Bread Wheat. *Plant Cell, Tissue and Organ Culture*, 77, 245-250.
- Peng, B., Li, Y., Wang, Y., Liu, C., Liu, Z., Zhang, Y., Tan, W., Wang, D., Shi, Y., Sun, B., et al. (2013). Correlations and comparisons of quantitative trait loci with family per se and testcross performance for grain yield and related traits in maize. *Theor. Appl. Genet.* 126, 773–789.
- Pereyra-Irujo, G.A., Velázquez, L., Lechner, L., and Aguirrezábal, L.A.N. (2008). Genetic variability for leaf growth rate and duration under water deficit in sunflower: analysis of responses at cell, organ, and plant level. *J. Exp. Bot.* 59, 2221–2232.
- Piepho, H.P., Möhring, J., Melchinger, A.E., and Büchse, A. (2007). BLUP for phenotypic selection in plant breeding and variety testing. *Euphytica* 161, 209–228.
- Poormohammad Kiani, S., Maury, P., Nouri, L., Ykhlef, N., Grieu, P., and Sarrafi, A. (2009). QTL analysis of yield-related traits in sunflower under different water treatments. *Plant Breeding* 128, 363–373.
- Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* 38, 904–909.
- Pritchard, J.K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* 155, 945–959.
- Pustovoit, V.S. (1964). Conclusions of work on the selection and seed production of sunflowers (in Russian). *Agrobiology*, 5, 662-697
- Pustovoit, V.S (1967). Handbook of selection and seed growing of oil plants. Kolos, Moscow
- Rauf, S., Sadaqat, H.A., and Khan, I.A. (2008). Effect of moisture regimes on combining ability variations of seedling traits in sunflower (*Helianthus annuus* L.). *Canadian Journal of Plant Science* 88, 323–329.
- Rengel, D., Arribat, S., Pierre Maury, P., Martin-Magniette, M.L., Hourlier, T., Laporte, M., Varès, D., Carrère, C., Grieu, P., Balzergue, S., Gouzy, J., Vincourt, P., Langlade, N.B (2012) A Gene-Phenotype Network Based on Genetic Variability for Drought Responses Reveals Key Physiological Processes in Controlled and Natural Environments. *PLoS ONE* 7(10): e45249 DOI
- Reymond M, Muller B, Leonardi A, Charcosset A and Tardieu F (2003). Combining quantitative trait loci analysis and an ecophysiological model to analyze the genetic variability of the responses of maize leaf growth to temperature and water deficit. *Plant Physiology* 131, 664–675.
- Ribaut, J.M., Ragot, M. (2007). Marker-assisted selection to improve drought adaptation in maize: the backcross approach, perspectives, limitations and alternatives. *Journal of Experimental Botany* 58, 351–360.
- Ribaut, J-M, Hoisington, D, Banziger, M, Setter, T, and Edmeades, G. (2004). Genetic dissection of drought tolerance in maize: a case study. In: *Physiology and biotechnology integration for plant breeding* (Nguyen HT and Blum A, eds). Marcel Dekker, Inc, New York, USA, pp 571–609.



- Rieseberg, L.H. (1995). The role of hybridization in evolution: old wine in new skins. *American Journal of Botany* 82, 944–953.
- Roche, J., Hewezi, T., Bouniols, A., and Gentzbittel, L. (2007). Transcriptional profiles of primary metabolism and signal transduction-related genes in response to water stress in field-grown sunflower genotypes using a thematic cDNA microarray. *Planta* 226, 601–617.
- Rondanini, D., R. Savin and A.J. Hall, (2003). Dynamics of fruit growth and oil quality of sunflower (*Helianthus annuus* L.) exposed to brief intervals of high temperature during grain filling. *Field Crop Res.*, 83, 79-90.
- Sabetta, W., Alba, V., Blanco, A., and Montemurro, C. (2011). sunTILL: a TILLING resource for gene function analysis in sunflower. *Plant Methods* 7, 20.
- Salekdeh, G.H., Reynolds, M., Bennett, J., and Boyer, J. (2009). Conceptual framework for drought phenotyping during molecular breeding. *Trends Plant Sci.* 14, 488–496.
- Self, S.G., and Liang, K.-Y. (1987). Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests Under Nonstandard Conditions. *Journal of the American Statistical Association* 82, 605.
- Segura, V., Vilhjálmsson, B.J., Platt, A., Korte, A., Seren, Ü., Long, Q., and Nordborg, M. (2012). An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat Genet* 44, 825–830
- Setter, T.L. (2012). Analysis of constituents for phenotyping drought tolerance in crop improvement. *Front Physiol* 3, 180.
- Shindo, T., Misas-Villamil, J.C., Hörger, A.C., Song, J., and van der Hoorn, R.A.L. (2012). A Role in Immunity for Arabidopsis Cysteine Protease RD21, the Ortholog of the Tomato Immune Protease C14. *PLoS ONE* 7, e29317.
- Shinozaki, K., Yamaguchi-Shinozaki, K. (2007) Gene networks involved in drought stress response and tolerance. *J Exp Bot* 58, 221–227
- Skoric, D. (2009). Sunflower breeding for resistance to abiotic stresses. *Helia* 32, 1–15.
- Smith AB, Cullis BR and Thompson R (2005). The analysis of crop cultivar breeding and evaluation trials: an overview of current mixed model approaches. *Journal of Agricultural Science* 143, 449–462.
- Snow, A. A., P. Moran-Palma, L. H. Rieseberg, A. Wszelaki, and G. J. Seiler (1998). Fecundity, phenology, and seed dormancy of F1 wild-crop hybrids in sunflower (*Helianthus annuus*, Asteraceae). *American Journal of Botany* 85, 794–801.
- Soriano, M.A., Orgaz, F., Villalobos, F.J., and Fereres, E. (2004). Efficiency of water use of early plantings of sunflower. *European Journal of Agronomy* 21, 465–476.
- Steele, K.A., Price, A.H., Shashidhar, H.E., and Witcombe, J.R. (2006). Marker-assisted selection to introgress rice QTLs controlling root traits into an Indian upland rice variety. *Theor. Appl. Genet.* 112, 208–221.
- Stoenescu FGenetics. In: Vranceau AV, editor. Floarea-soarelui. Bucuresti (Romania): Academiei Republicii Socialiste; 1974. p. 92-125.



- Sun, G., Zhu, C., Kramer, M.H., Yang, S.S., Song, W., Piepho, H.P., and Yu, J. (2010). Variation explained in mixed-model association mapping. *Heredity* 105, 333–340.
- Tang S, Yu JK., Slabaugh MB, Shintani DK, Knapp SJ (2002) Simple sequence repeat map of the sunflower genome. *Theor Appl Genet* 105,1124-1136
- Tardieu, F. (2011). Any trait or trait-related allele can confer drought tolerance: just design the right drought scenario. *J. Exp. Bot.* err269.
- Teixeira, J., and 36 co-authors (2011) Tropical and sub-tropical cloud transitions in weather and climate prediction models: the GCSS/WGNE Pacific Cross-section Intercomparison (GPCI). *Journal of Climate*, 24, 5223-5256
- Tenaillon, M. I., Sawkins, M.C., Long ,A.D., Gaut, R.L., Doebley, J.F., and Gaut B.S. (2001). Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). *Proc. Natl. Acad. Sci. USA*. 98, 9161-9166.
- Tersac, M., Blanchard, P., Brunel, D., Vincourt, P. (1994). Relations between heterosis and enzymatic polymorphism in populations of cultivated sunflowers (*Helianthus annuus* L.). *Theor Appl Genet* 88, 49–55
- Tersac M, Vares D, Vincourt P (1993) Combining groups in cultivated sunflower populations (*Helianthus annuus* L.) and their relationships
- Thornsberry, J.M., Goodman, M.M., Doebley, J., Kresovich, S., Nielsen, D., Buckler, E.S., and others (2001). Dwarf8 polymorphisms associate with variation in flowering time. *Nature Genetics* 28, 286–289.
- Tian, D., Araki, H., Stahl, E., Bergelson, J., and Kreitman, M. (2002). Signature of balancing selection in *Arabidopsis*. *Proc Natl Acad Sci U S A* 99, 11525–11530.
- Tian, F., Bradbury, P.J., Brown, P.J., Hung, H., Sun, Q., Flint-Garcia, S., Rocheford, T.R., McMullen, M.D., Holland, J.B., and Buckler, E.S. (2011). Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nature Genetics* 43, 159–162.
- Tuberosa R, Salvi S, Sanguineti MC, Landi P, Maccaferri M and Conti S (2002a). Mapping QTLs regulating morpho-physiological traits and yield: Case studies, shortcomings and perspectives in drought-stressed maize. *Annals of Botany* 89, 941–963.
- Van Inghelandt, D., Melchinger, A.E., Lebreton, C., and Stich, B. (2010). Population structure and genetic diversity in a commercial maize breeding program assessed with SSR and SNP markers. *Theor Appl Genet* 120, 1289–1299.
- Varshney, R.K., Paulo, M.J., Grando, S., van Eeuwijk, F.A., Keizer, L.C.P., Guo, P., Ceccarelli, S., Kilian, A., Baum, M., and Graner, A. (2012). Genome wide association analyses for drought tolerance related traits in barley (*Hordeum vulgare* L.). *Field Crops Research* 126, 171–180.
- Vear, F., Muller, M.H. (2011) Progrès variétal chez le tournesol : l'apport des ressources génétiques au sein du genre *Helianthus*. *Innovations Agronomiques* 14, 139-150
- Vear, F., Serre, F., Jouan-Dufournel, I., Bert, P.F., Roche, S., Walser, P., Tourvieille de Labrouhe, D., and Vincourt, P. (2008). Inheritance of quantitative resistance to downy mildew (*Plasmopara halstedii*) in sunflower (*Helianthus annuus* L.). *Euphytica* 164, 561–570.



- Venuprasad R, Lafitte HR, Atlin GN (2007) Response to direct selection for grain yield under drought stress in rice. *Crop Sci* 47, 285–293
- Villalobos, F., Hall, A., Ritchie, J., Orgaz, F. (1996). OILCROP-SUN: a development, growth and yield model of sunflower crop. *Agronomy Journal* 88, 403–415.
- Vincourt P., Vear F. (2009). Les varieties hybrides chez le tournesol. *Le Sélectionneur Français* 60, 31-38
- Vincourt, P., As Sadi, F., Bordat, A., Langlade, N., Gouzy, J., Pouilly, N., Lippi, Y., Serre, F., Godiard L., Tourvieille de Labrouhe, D., Vear, F. (2012) Consensus mapping of major resistance genes and independent QTL for quantitative resistance to sunflower downy mildew *Theor Appl Genet* 5, 909–920
- Visioni, A., Tondelli, A., Francia, E., Pswarayi, A., Malosetti, M., Russell, J., Thomas, W., Waugh, R., Pecchioni, N., Romagosa, I., et al. (2013). Genome-wide association mapping of frost tolerance in barley (*Hordeum vulgare* L.). *BMC Genomics* 14, 424.
- Vranceanu, A.V. (1974). Sunflower. Academy of Romanian Socialist Republic, Bucharest
- Wallace, J.G., Larsson, S., Buckler, E.S. (2013). Entering the Second Century of Maize Quantitative Genetics. *Heredity*
- Wang, W., Vinocur, B., Altman, A., (2003). Plant response to drought, salinity and extreme temperatures: toward genetic engineering for stress tolerance. *Planta* 218, 1–14.
- Wang, J., McClean, P.E., Lee, R., Goos, R.J., and Helms, T. (2008). Association mapping of iron deficiency chlorosis loci in soybean (*Glycine max* L. Merr.) advanced breeding lines. *Theoretical and Applied Genetics* 116, 777–787.
- Whitney, K.D., Randell, R.A., and Rieseberg, L.H. (2010). Adaptive introgression of abiotic tolerance traits in the sunflower *Helianthus annuus*. *New Phytologist* 187, 230–239.
- Wills, D.M., and Burke, J.M. (2007). Quantitative Trait Locus Analysis of the Early Domestication of Sunflower. *Genetics* 176, 2589–2599.
- Xue, Y., Warburton, M.L., Sawkins, M., Zhang, X., Setter, T., Xu, Y., Grudloyma, P., Gethi, J., Ribaut, J.-M., Li, W., et al. (2013). Genome-wide association analysis for nine agronomic traits in maize under well-watered and water-stressed conditions. *Theor Appl Genet* 126, 2587–2596.
- Yadav, R.S., Sehgal, D., Vadez, V. (2011). Using genetic mapping and genomics approaches in understanding and improving drought tolerance in pearl millet. *Journal of Experimental Botany* 62, 397–408.
- Yan, J., Shah, T., Warburton, M.L., Buckler, E.S., McMullen, M.D., and Crouch, J. (2009). Genetic Characterization and Linkage Disequilibrium Estimation of a Global Maize Collection Using SNP Markers. *America* 4.
- Yu, J., Pressoir, G., Briggs, W.H., Bi, I.V., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S., Nielsen, D.M., Holland, J.B., et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics* 38, 203–208.
- Yu, L., Chen, X., Wang, Z., Wang, S., Wang, Y., Zhu, Q., Li, S., and Xiang, C. (2013). Arabidopsis Enhanced Drought Tolerance1/HOMEODOMAIN GLABROUS11 Confers Drought Tolerance in Transgenic Rice without Yield Penalty. *Plant Physiology* 162, 1378–1391.

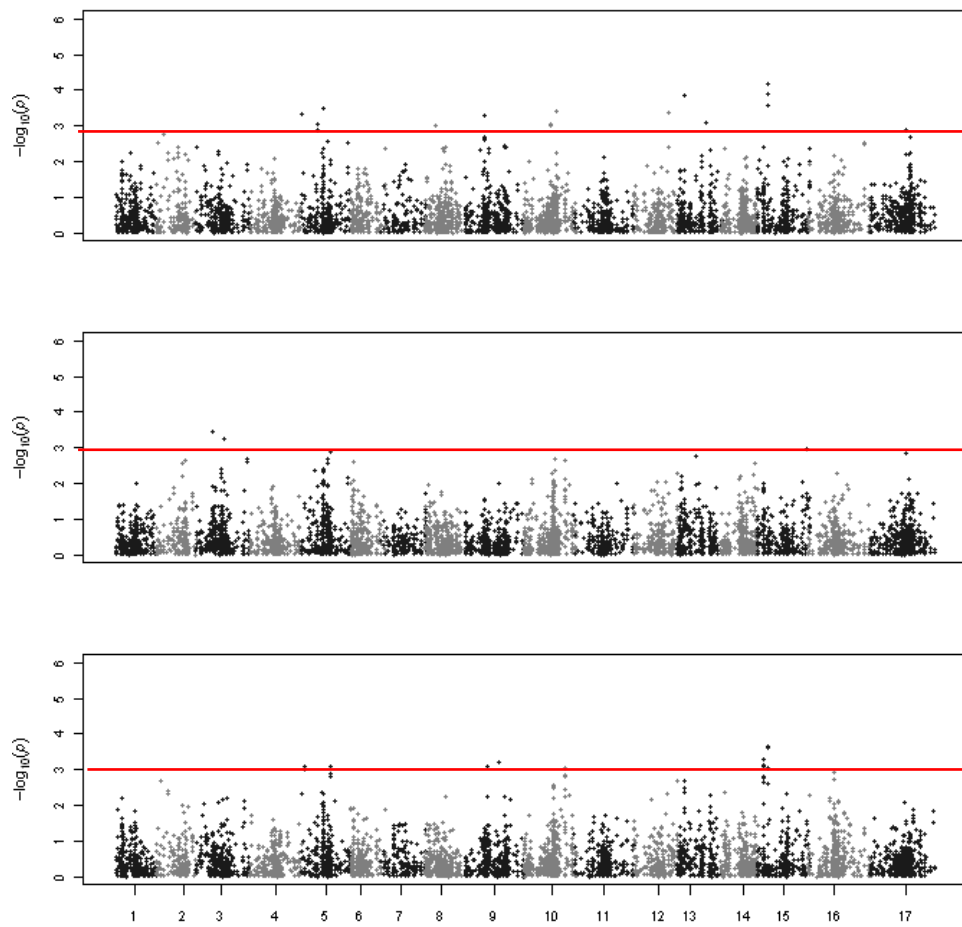




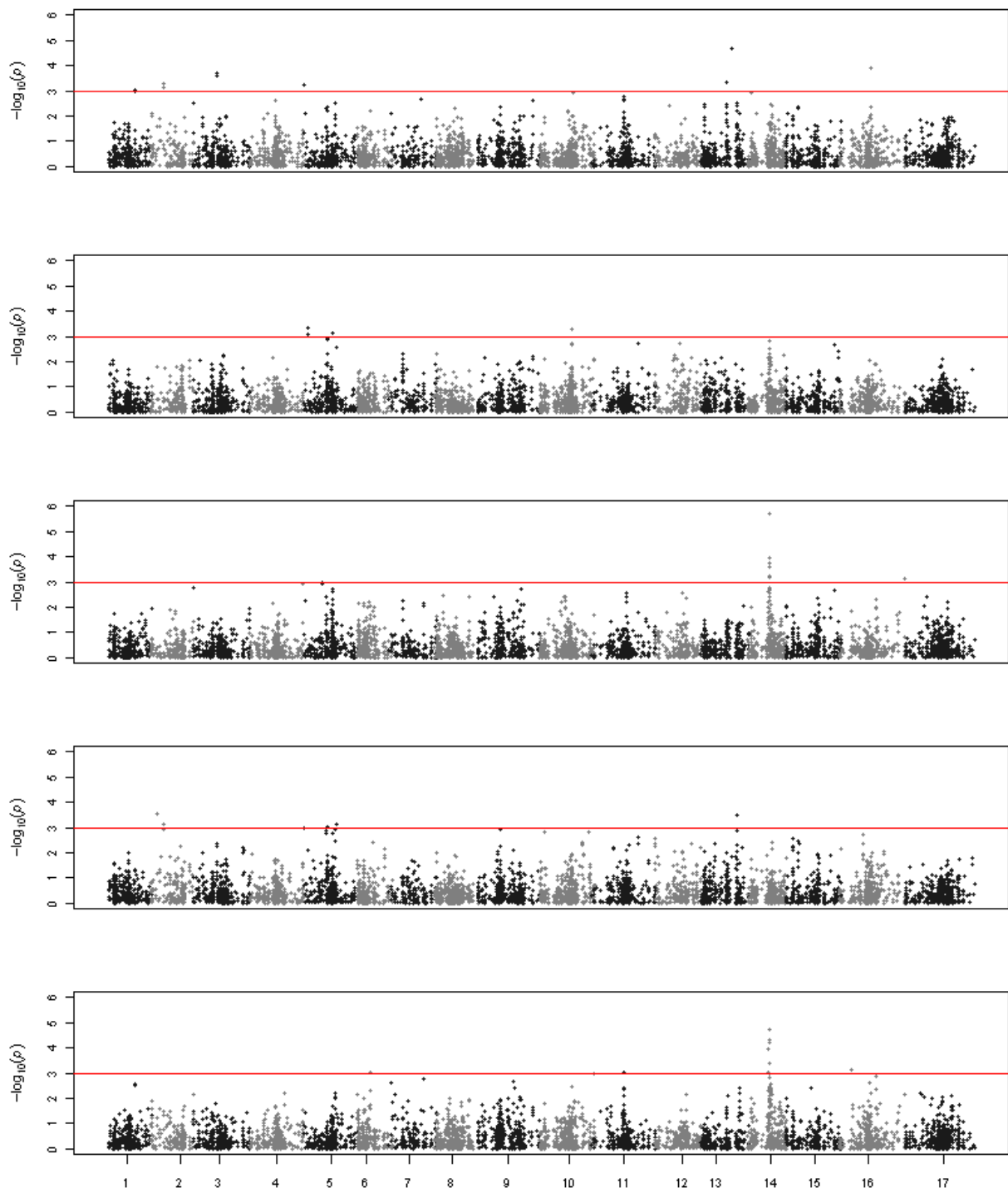
- Zhang, Y.X., Gentzbittel, L., Vear, F., and Nicolas, P. (1995). Assessment of inter- and intra-inbred line variability in sunflower (*Helianthus annuus*) by RFLPs. *Genome* 38, 1040–1048.
- Zhang, Z., Ersoz, E., Lai, C.-Q., Todhunter, R.J., Tiwari, H.K., Gore, M.A., Bradbury, P.J., Yu, J., Arnett, D.K., Ordovas, J.M., et al. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics* 42, 355–360.
- Zhao, K., Tung, C.-W., Eizenga, G.C., Wright, M.H., Ali, M.L., Price, A.H., Norton, G.J., Islam, M.R., Reynolds, A., Mezey, J., et al. (2011). Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*. *Nat Commun* 2, 467.
- Zhao, K., Aranzana, M.J., Kim, S., Lister, C., Shindo, C., Tang, C., Toomajian, C., Zheng, H., Dean, C., Marjoram, P., et al. (2007). An Arabidopsis example of association mapping in structured samples. *PLoS Genetics* 3, e4.
- Zheng, B.S., Gouis, J., Leflon, M., Rong, W.Y., Laperche, A., and Brancourt-Hulmel, M. (2010). Using probe genotypes to dissect QTL  $\times$  environment interactions for grain yield components in winter wheat. *Theoretical and Applied Genetics* 121, 1501–1517.
- Zhou, X., and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics* 44, 821–824.
- Zou, J., Jiang, C., Cao, Z., Li, R., Long, Y., Chen, S., and Meng, J. (2010). Association mapping of seed oil content in *Brassica napus* and comparison with quantitative trait loci identified from linkage mapping. *Genome* 53, 908–916.



# Annexes



**Figure S0 : Manhattan plot des variables synthétiques liées au rendement grains : modèle RDT =f(JS). De haut en bas : RDT\_moy, RDT\_TOLH, RDT\_SEL**



**Figure S0 bis : Manhattan plot des variables synthétiques liées au rendement en huile : modèle  $RDTH = f(JS + IFT)$ . De haut en bas: RDTH\_moy, RDTH\_TOLH, RDTH\_TOLT, RDTH\_SELH, RDTH\_SELT**

SNP	chromosome	position	caractère	environnement	p-values	R <sup>2</sup> <sub>LR</sub>	moyenne pheno	allele 0	allele 1	effet allelique 0 - 1
HS153121	LG01	25.5	Hauteur	AI09_I	3.67E-06	0.07	132.88	AA	TT	-11.66796627
HS153121	LG01	25.5	M3	CO08_NI	0.00017322	0.04	252.54	AA	TT	-2.702917133
HS153121	LG01	25.5	H2O	GA09_I	5.52E-06	0.06	4.64	AA	TT	-0.826124904
HS153121	LG01	25.5	Hauteur	CA10	3.98E-07	0.07	156.20	AA	TT	-12.82299517
HS151306	LG01	26.1	IF40	CA10	9.38E-08	0.09	0.51	AA	GG	0.520325375
HS141137	LG01	27.6	Hauteur	AI09_I	9.32E-07	0.08	132.88	CC	TT	-7.791356905
HS141137	LG01	27.6	H2O	AI09_NI	0.00032051	0.04	6.05	CC	TT	-1.015461346
HS147758	LG01	34.2	LAI_F1	CA10	4.68E-05	0.05	4.42	AA	GG	-0.135184602
HS083828	LG01	40	RDTH	AI08_I	1.73E-05	0.05	16.74	CC	TT	0.702231038
HS092129	LG01	40	RDTH	AI08_I	1.71E-05	0.05	16.74	AA	GG	-0.702336832
HS083828	LG01	40	RDTH	VE10	6.81E-06	0.06	16.15	CC	TT	0.796607053
HS092129	LG01	40	RDTH	VE10	6.57E-06	0.06	16.15	AA	GG	-0.793128252
HS128327	LG01	60.9	IF40	AI08_NI	9.33E-06	0.07	2.48	GG	TT	0.390559137
HS075253	LG02	30.8	LAI_F1	CA10	3.89E-05	0.06	4.42	CC	GG	-0.114653527
HS129134	LG02	45.3	H2O	AI09_NI	4.94E-05	0.05	6.05	AA	GG	-1.436722181
HS129134	LG02	45.3	IF40	CO08_NI	2.27E-06	0.07	1.58	AA	GG	-0.617501044
HS129134	LG02	45.3	H2O	VE09_I	0.00011796	0.04	6.10	AA	GG	-1.10278811
HS068866	LG02	46.3	M3	CO08_NI	1.60E-05	0.05	252.54	GG	TT	-2.050709011
HS147170	LG02	60.4	PMG	CO08_NI	6.75E-05	0.05	43.70	AA	GG	-1.959149686
HS105597	LG03	17.5	RDTH	VER09_NI	4.33E-05	0.05	15.56	CC	TT	0.464434795
HS061677	LG03	22.6	PMG	CO08_NI	5.48E-05	0.05	43.70	AA	GG	-1.972894873
HS062090	LG03	26.8	LAD	CO09_NI	5.30E-06	0.06	40.36	AA	GG	1.212302155
HS090132	LG03	26.8	LAD	CO09_NI	5.22E-06	0.06	40.36	CC	TT	1.231620222
HS095348	LG03	26.8	LAD	CO09_NI	7.40E-07	0.07	40.36	GG	TT	-1.331744397
HS054380	LG03	26.8	LAD	CO09_NI	2.58E-05	0.05	40.36	AA	GG	1.086548569
HS142400	LG03	26.8	LAD	CO09_NI	1.70E-05	0.06	40.36	CC	TT	-1.140501495
HS068334	LG03	26.8	LAD	CO09_NI	2.68E-05	0.05	40.36	AA	GG	1.098371372
HS082361	LG03	54.1	H2O	CO09_I	3.25E-06	0.06	6.34	AA	CC	0.423635868
HS058930	LG04	11	PMG	VE09_I	1.39E-05	0.06	51.64	AA	TT	-1.726516382
HS083365	LG04	27.9	RDTH	GA09_I	8.99E-06	0.06	12.32	CC	GG	0.554471699
HS083365	LG04	27.9	Nbgrains	GA09_I	3.94E-06	0.06	79107.84	CC	GG	3024.433055
HS151336	LG04	34	F1	AI09_NI	0.00011398	0.04	189.09	AA	GG	1.913215916
HS151336	LG04	34	H2O	VE09_I	9.39E-05	0.04	6.10	AA	GG	1.011178077
HS058216	LG04	40.7	H2O	CO08_I	6.94E-05	0.05	6.57	CC	TT	0.957525335
HS058216	LG04	40.7	Hauteur	CA10	1.60E-05	0.05	156.20	CC	TT	10.8596516
HS063772	LG04	42.5	IF40	VER09_NI	8.39E-07	0.08	4.03	AA	GG	0.452955897
HS063772	LG04	42.5	Senescevol	VER09_NI	2.87E-05	0.05	155.56	AA	GG	9.146847254
HS103244	LG04	42.7	H2O	AI08_I	1.75E-05	0.05	7.54	CC	TT	1.438699744
HS052343	LG04	71.6	F1	AI08_NI	1.38E-05	0.06	201.03	CC	GG	0.909007787
HS147113	LG05	34.3	M3	CO09_I	8.05E-06	0.06	247.20	AA	GG	1.78171312
HS138568	LG05	34.3	H2O	VE09_I	0.00012162	0.04	6.10	CC	GG	-1.189557907
HS138568	LG05	34.3	H2O	VER09_NI	2.58E-05	0.05	5.62	CC	GG	-1.484426357
HS147125	LG05	34.3	RDTH	SE10	1.42E-05	0.05	10.88	CC	TT	-0.762115856
HS147124	LG05	34.3	RDTH	SE10	6.71E-06	0.06	10.88	AA	CC	0.836465504
HS073823	LG05	34.5	H2O	AI09_NI	0.00014996	0.04	6.05	CC	TT	1.324380886
HS134850	LG05	40.6	M3	CO08_NI	0.00013559	0.04	252.54	AA	CC	1.941219751
HS107108	LG05	41.3	F1	VE09_I	6.29E-05	0.05	190.12	AA	GG	-1.72998414
HS157386	LG06	7.4	Hauteur	GA09_NI	1.18E-05	0.05	144.14	GG	TT	2.365382546
HS137161	LG06	10.8	Senescevol	VER09_NI	7.83E-05	0.05	155.56	AA	GG	9.20682041
HS145597	LG07	14.5	H2O	CHA10	6.61E-05	0.05	10.52	CC	TT	1.516085925
HS075061	LG08	2.8	H2O	AI09_I	1.84E-05	0.05	6.86	AA	GG	-1.078038259
HS131212	LG08	2.8	H2O	AI09_I	8.11E-05	0.05	6.86	AA	GG	-0.992319329
HS148271	LG08	2.8	H2O	AI09_I	2.99E-05	0.05	6.86	CC	TT	1.128591093
HS150536	LG08	2.8	H2O	AI09_I	2.99E-05	0.05	6.86	CC	TT	1.128591093
HS148271	LG08	2.8	H2O	AI09_NI	0.0001491	0.04	6.05	CC	TT	0.878989179
HS150536	LG08	2.8	H2O	AI09_NI	0.0001491	0.04	6.05	CC	TT	0.878989179
HS093777	LG08	2.8	H2O	CO08_I	2.76E-05	0.05	6.57	CC	TT	-0.841308439
HS148271	LG08	2.8	H2O	VE09_I	5.52E-05	0.05	6.10	CC	TT	0.737726999
HS150536	LG08	2.8	H2O	VE09_I	5.52E-05	0.05	6.10	CC	TT	0.737726999
HS156573	LG08	32.2	H2O	VE09_I	1.69E-06	0.05	6.10	AA	GG	1.367077932
HS070642	LG08	32.3	RDT	VE10	6.80E-05	0.04	34.40	AA	GG	1.82248109
HS117040	LG09	30	F1	AI09_I	1.67E-06	0.06	189.84	CC	TT	1.876935317
HS107618	LG09	30	F1	AI09_I	1.67E-06	0.06	189.84	CC	TT	1.876935317
HS117040	LG09	30	F1	AI09_NI	5.16E-06	0.06	189.09	CC	TT	1.61184707
HS107618	LG09	30	F1	AI09_NI	5.16E-06	0.06	189.09	CC	TT	1.61184707
HS117040	LG09	30	IF20	AI09_NI	7.33E-06	0.06	7.41	CC	TT	0.220029708

SNP	chromosome	position	caractère	environnement	p-values	R <sup>2</sup> <sub>LR</sub>	moyenne pheno	allele 0	allele 1	effet allelique 0 - 1
HS107618	LG09	30	IF20	AI09_NI	7.33E-06	0.06	7.41	CC	TT	0.220029708
HS117040	LG09	30	IF30	AI09_NI	2.19E-05	0.05	5.94	CC	TT	0.346445637
HS107618	LG09	30	IF30	AI09_NI	2.19E-05	0.05	5.94	CC	TT	0.346445637
HS117040	LG09	30	IF40	AI09_NI	1.20E-06	0.07	3.33	CC	TT	0.511667162
HS107618	LG09	30	IF40	AI09_NI	1.20E-06	0.07	3.33	CC	TT	0.511667162
HS117040	LG09	30	Senescevol	AI09_NI	7.34E-07	0.07	123.82	CC	TT	7.670790164
HS107618	LG09	30	Senescevol	AI09_NI	7.34E-07	0.07	123.82	CC	TT	7.670790164
HS117040	LG09	30	F1	CO08_I	7.19E-05	0.05	212.89	CC	TT	1.45840332
HS107618	LG09	30	F1	CO08_I	7.19E-05	0.05	212.89	CC	TT	1.45840332
HS117040	LG09	30	M0	CO08_I	3.94E-05	0.05	217.56	CC	TT	1.588984285
HS107618	LG09	30	M0	CO08_I	3.94E-05	0.05	217.56	CC	TT	1.588984285
HS117040	LG09	30	M0	CO08_NI	3.04E-05	0.05	217.83	CC	TT	1.677221073
HS107618	LG09	30	M0	CO08_NI	3.04E-05	0.05	217.83	CC	TT	1.677221073
HS117040	LG09	30	F1M0	CO08_NI	4.06E-05	0.05	4.45	CC	TT	0.201658756
HS107618	LG09	30	F1M0	CO08_NI	4.06E-05	0.05	4.45	CC	TT	0.201658756
HS117040	LG09	30	F1	VE09_I	2.29E-05	0.05	190.12	CC	TT	1.501456983
HS107618	LG09	30	F1	VE09_I	2.29E-05	0.05	190.12	CC	TT	1.501456983
HS117040	LG09	30	IF30	VE09_I	3.95E-05	0.05	6.15	CC	TT	0.232910036
HS107618	LG09	30	IF30	VE09_I	3.95E-05	0.05	6.15	CC	TT	0.232910036
HS117040	LG09	30	Senescevol	VE09_I	3.90E-05	0.05	167.47	CC	TT	9.886374274
HS107618	LG09	30	Senescevol	VE09_I	3.90E-05	0.05	167.47	CC	TT	9.886374274
HS117040	LG09	30	IF30	VER09_NI	3.16E-05	0.05	5.64	CC	TT	0.228147372
HS107618	LG09	30	IF30	VER09_NI	3.16E-05	0.05	5.64	CC	TT	0.228147372
HS117040	LG09	30	Senescevol	VER09_NI	1.43E-05	0.05	155.56	CC	TT	7.433272001
HS107618	LG09	30	Senescevol	VER09_NI	1.43E-05	0.05	155.56	CC	TT	7.433272001
HS117040	LG09	30	RDT	VE10	8.56E-05	0.05	34.40	CC	TT	1.424979641
HS107618	LG09	30	RDT	VE10	8.56E-05	0.05	34.40	CC	TT	1.424979641
HS117598	LG09	30.3	Hauteur	AI09_I	1.18E-06	0.06	132.88	CC	TT	-7.357184619
HS117598	LG09	30.3	H2O	AI09_I	1.95E-06	0.07	6.86	CC	TT	-1.388340533
HS106242	LG09	30.3	H2O	AI09_I	0.00016102	0.04	6.86	AA	GG	1.014014398
HS090401	LG09	30.3	F1	AI09_NI	1.26E-06	0.06	189.09	AA	TT	1.6527137
HS106242	LG09	30.3	F1	AI09_NI	2.89E-07	0.07	189.09	AA	GG	1.731160016
HS090401	LG09	30.3	IF20	AI09_NI	2.34E-06	0.06	7.41	AA	TT	0.231826962
HS106242	LG09	30.3	IF20	AI09_NI	4.07E-06	0.06	7.41	AA	GG	0.225832689
HS117598	LG09	30.3	IF30	AI09_NI	1.78E-05	0.05	5.94	CC	TT	-0.391953981
HS090401	LG09	30.3	IF30	AI09_NI	1.23E-07	0.08	5.94	AA	TT	0.422411458
HS106242	LG09	30.3	IF30	AI09_NI	9.71E-08	0.08	5.94	AA	GG	0.424024802
HS117598	LG09	30.3	IF40	AI09_NI	1.01E-06	0.07	3.33	CC	TT	-0.562678943
HS090401	LG09	30.3	IF40	AI09_NI	4.10E-08	0.09	3.33	AA	TT	0.577481543
HS106242	LG09	30.3	IF40	AI09_NI	1.99E-08	0.09	3.33	AA	GG	0.586590158
HS117598	LG09	30.3	Senescevol	AI09_NI	1.69E-06	0.06	123.82	CC	TT	-8.26749742
HS117598	LG09	30.3	F1	CO08_I	5.33E-07	0.07	212.89	CC	TT	-2.009376801
HS090401	LG09	30.3	F1	CO08_I	2.71E-05	0.05	212.89	AA	TT	1.537841165
HS106242	LG09	30.3	F1	CO08_I	4.94E-06	0.06	212.89	AA	GG	1.658040086
HS117598	LG09	30.3	M0	CO08_I	7.67E-08	0.08	217.56	CC	TT	-2.23148028
HS090401	LG09	30.3	M0	CO08_I	1.56E-05	0.05	217.56	AA	TT	1.677669238
HS106242	LG09	30.3	M0	CO08_I	2.43E-06	0.06	217.56	AA	GG	1.813350402
HS117598	LG09	30.3	F1	CO08_NI	1.20E-06	0.07	213.42	CC	TT	-2.064824871
HS106242	LG09	30.3	F1	CO08_NI	2.40E-05	0.05	213.42	AA	GG	1.62533546
HS090401	LG09	30.3	M0	CO08_NI	1.71E-06	0.07	217.83	AA	TT	1.900720518
HS106242	LG09	30.3	M0	CO08_NI	1.09E-06	0.07	217.83	AA	GG	1.923645651
HS117598	LG09	30.3	M3	CO08_NI	3.49E-05	0.05	252.54	CC	TT	-1.584205802
HS117598	LG09	30.3	H2O	CO08_NI	1.99E-06	0.07	13.45	CC	TT	-1.340118215
HS090401	LG09	30.3	F1M0	CO08_NI	4.86E-05	0.05	4.45	AA	TT	0.203653081
HS090401	LG09	30.3	F1	VE09_I	7.72E-07	0.07	190.12	AA	TT	1.708563829
HS106242	LG09	30.3	F1	VE09_I	3.12E-07	0.07	190.12	AA	GG	1.751021559
HS090401	LG09	30.3	IF40	VE09_I	2.12E-05	0.05	4.31	AA	TT	0.54078064
HS106242	LG09	30.3	IF40	VE09_I	1.88E-05	0.05	4.31	AA	GG	0.540599941
HS090401	LG09	30.3	Senescevol	VE09_I	1.30E-05	0.05	167.47	AA	TT	10.32067922
HS106242	LG09	30.3	Senescevol	VE09_I	1.45E-05	0.05	167.47	AA	GG	10.19784767
HS117598	LG09	30.3	F1	VER09_NI	1.77E-07	0.09	191.68	CC	TT	-2.032690622
HS090401	LG09	30.3	F1	VER09_NI	1.50E-05	0.06	191.68	AA	TT	1.520124076
HS106242	LG09	30.3	F1	VER09_NI	1.83E-06	0.07	191.68	AA	GG	1.659609243
HS106242	LG09	30.3	IF30	VER09_NI	3.91E-05	0.05	5.64	AA	GG	0.22408042
HS090401	LG09	30.3	IF40	VER09_NI	4.21E-05	0.06	4.03	AA	TT	0.29098242

HS106242	LG09	30.3	IF40	VER09_NI	1.47E-05	0.06	4.03	AA	GG	0.306138753
HS117598	LG09	30.3	Senescevol	VER09_NI	7.74E-06	0.06	155.56	CC	TT	-8.762782206
HS090401	LG09	30.3	Senescevol	VER09_NI	2.10E-06	0.06	155.56	AA	TT	8.125922086
HS106242	LG09	30.3	Senescevol	VER09_NI	6.86E-07	0.07	155.56	AA	GG	8.43768892
HS117598	LG09	30.3	F1	CA10	3.91E-05	0.05	196.47	CC	TT	-1.891861789
HS090401	LG09	30.3	F1	CA10	0.00011729	0.04	196.47	AA	TT	1.556875562
HS106242	LG09	30.3	F1	CA10	6.24E-05	0.04	196.47	AA	GG	1.595561956
HS117598	LG09	30.3	M0	CA10	1.50E-05	0.06	211.72	CC	TT	-1.549396441
HS117598	LG09	30.3	H2O	CHA10	3.60E-05	0.05	10.52	CC	TT	-1.323992665
HS090401	LG09	30.3	H2O	CHA10	0.00012909	0.04	10.52	AA	TT	1.104060044
HS090401	LG09	30.3	RDT	VE10	1.06E-05	0.06	34.40	AA	TT	1.635687454
HS106242	LG09	30.3	RDT	VE10	2.10E-06	0.06	34.40	AA	GG	1.752611848
HS160723	LG09	32	F1	AI09_I	3.27E-06	0.06	189.84	CC	TT	-1.875640005
HS160743	LG09	32	F1	AI09_I	8.64E-07	0.07	189.84	AA	TT	1.932593519
HS160723	LG09	32	F1	AI09_NI	1.58E-05	0.05	189.09	CC	TT	-1.540637074
HS160743	LG09	32	F1	AI09_NI	6.12E-06	0.06	189.09	AA	TT	1.593202731
HS160723	LG09	32	IF20	AI09_NI	4.51E-05	0.05	7.41	CC	TT	-0.219025221
HS160743	LG09	32	IF20	AI09_NI	6.04E-06	0.06	7.41	AA	TT	0.225891701
HS160743	LG09	32	IF30	AI09_NI	7.44E-05	0.04	5.94	AA	TT	0.325683656
HS160743	LG09	32	IF40	AI09_NI	7.63E-06	0.06	3.33	AA	TT	0.47811041
HS160723	LG09	32	Senescevol	AI09_NI	5.91E-05	0.05	123.82	CC	TT	-6.552944473
HS160743	LG09	32	Senescevol	AI09_NI	2.56E-06	0.06	123.82	AA	TT	7.354278265
HS160723	LG09	32	M0	CO08_I	6.13E-05	0.05	217.56	CC	TT	-1.599227066
HS160743	LG09	32	M0	CO08_I	4.02E-05	0.05	217.56	AA	TT	1.574010683
HS160743	LG09	32	M0	CO08_NI	8.96E-05	0.04	217.83	AA	TT	1.542511379
HS160743	LG09	32	F1M0	CO08_NI	3.55E-05	0.05	4.45	AA	TT	0.205677914
HS164530	LG09	32	H2O	CO09_I	4.39E-06	0.06	6.34	CC	TT	0.469858387
HS164549	LG09	32	H2O	CO09_I	2.44E-05	0.05	6.34	CC	TT	-0.44789682
HS160723	LG09	32	F1	VE09_I	1.53E-06	0.07	190.12	CC	TT	-1.6764267
HS160743	LG09	32	F1	VE09_I	5.92E-06	0.06	190.12	AA	TT	1.616382826
HS160743	LG09	32	Senescevol	VE09_I	8.05E-05	0.04	167.47	AA	TT	9.595324721
HS160743	LG09	32	Senescevol	VER09_NI	5.12E-05	0.05	155.56	AA	TT	7.077042794
HS097368	LG09	64.8	H2O	AI09_NI	0.00019413	0.04	6.05	AA	GG	0.947586577
HS066491	LG10	34.5	M3	CO09_NI	9.53E-06	0.06	242.22	CC	TT	0.463935716
HS143048	LG10	43.9	F1	CO09_NI	4.53E-05	0.05	204.40	GG	TT	1.581538159
HS148756	LG10	43.9	F1	CO09_NI	5.84E-05	0.05	204.40	CC	TT	1.505799535
HS148757	LG10	43.9	F1	CO09_NI	4.53E-05	0.05	204.40	CC	TT	1.581538159
HS146707	LG10	45.6	RDTH	AI09_I	7.32E-06	0.06	20.31	CC	TT	-1.35052856
HS052625	LG10	45.8	RDTH	AI09_I	2.48E-05	0.05	20.31	AA	GG	0.893442176
HS159983	LG10	45.8	RDTH	AI09_I	9.58E-06	0.06	20.31	CC	GG	-1.31916247
HS097746	LG10	48.4	H2O	VE10	1.46E-05	0.06	7.72	AA	CC	-1.005841349
HS160400	LG10	51.2	RDT	AI08_NI	9.75E-06	0.06	28.07	GG	TT	-1.129160944
HS132684	LG10	51.7	RDT	AI08_NI	7.48E-08	0.08	28.07	CC	TT	-1.262708304
HS155342	LG10	76.8	IF20	CO08_NI	9.29E-05	0.05	4.94	CC	TT	-0.1391816
HS160632	LG11	26	Huille	CO09_I	2.88E-06	0.06	50.46	CC	TT	2.001822082
HS060364	LG11	48.2	Senescevol	VER09_NI	7.39E-05	0.04	155.56	CC	TT	8.062300136
HS144665	LG11	48.2	Senescevol	VER09_NI	7.39E-05	0.04	155.56	AA	GG	-8.062300136
HS053948	LG12	20.7	F1	CO09_NI	5.98E-05	0.05	204.40	AA	CC	0.900776362
HS147136	LG12	26.6	H2O	AI09_NI	0.00028143	0.04	6.05	CC	TT	-1.047269342
HS147136	LG12	26.6	M0	CO08_NI	6.34E-06	0.06	217.83	CC	TT	-2.498228792
HS147136	LG12	26.6	M3	CO08_NI	0.00016467	0.04	252.54	CC	TT	-1.954394941
HS147136	LG12	26.6	H2O	VE09_I	8.68E-05	0.05	6.10	CC	TT	-0.90333356
HS067550	LG12	38.2	M3	CO09_NI	5.87E-06	0.06	242.22	AA	TT	-0.428867691
HS123698	LG12	38.2	M3	CO09_NI	1.73E-06	0.07	242.22	AA	GG	0.401000655
HS149177	LG12	38.2	M3	CO09_NI	5.44E-06	0.06	242.22	AA	TT	-0.355351522
HS159700	LG12	38.2	M3	CO09_NI	1.07E-05	0.06	242.22	CC	TT	0.378316365
HS086480	LG12	44.3	Hauteur	GA09_NI	2.85E-05	0.05	144.14	GG	TT	-1.38725142
HS084642	LG12	56.8	H2O	AI09_I	1.19E-05	0.06	6.86	CC	GG	-2.1544061
HS084642	LG12	56.8	H2O	AI09_NI	0.00022025	0.04	6.05	CC	GG	-1.612353623
HS084642	LG12	56.8	M3	CO08_I	6.33E-05	0.05	254.38	CC	GG	-2.125265774
HS084642	LG12	56.8	H2O	GA09_NI	7.85E-06	0.07	8.61	CC	GG	-1.532686273
HS084642	LG12	56.8	H2O	VE09_I	0.00019054	0.04	6.10	CC	GG	-1.267253487
HS084642	LG12	56.8	RDT	VE10	9.12E-05	0.05	34.40	CC	GG	-2.706521479
HS077775	LG12	69.2	M3	VE10	2.58E-05	0.05	256.46	CC	TT	-2.882086687
HS077775	LG12	69.2	M0M3	VE10	1.33E-07	0.08	43.15	CC	TT	-2.943478432
HS141780	LG13	8.7	H2O	VE09_I	0.00012987	0.04	6.10	AA	CC	0.8308517
HS065348	LG13	38.3	IF20	CO08_NI	2.14E-05	0.06	4.94	GG	TT	0.090090488
HS069925	LG13	40	IF20	CO08_NI	0.0001265	0.05	4.94	AA	GG	0.083799275

HS067522	LG13	40	IF20	CO08_NI	1.56E-05	0.06	4.94	GG	TT	0.092369918
HS069925	LG13	40	Huile	VE09_I	3.98E-05	0.05	53.23	AA	GG	0.82730505
HS067522	LG13	40	Huile	VE09_I	5.78E-05	0.05	53.23	GG	TT	0.80567183
HS154746	LG13	40	Huile	VE09_I	2.95E-05	0.05	53.23	AA	GG	0.856866103
HS104806	LG13	53.3	RDTH	VER09_NI	2.08E-05	0.05	15.56	AA	GG	-0.525352644
HS056657	LG13	53.3	RDTH	VER09_NI	2.87E-05	0.05	15.56	CC	TT	0.501316261
HS099047	LG13	58.9	MOM3	CO08_I	8.71E-06	0.05	36.84	AA	TT	1.267130278
HS146281	LG13	60.8	LAI_F1	CA10	2.29E-05	0.06	4.42	AA	CC	0.195661565
HS057338	LG14	27	Senescevol	VE09_I	9.53E-05	0.05	167.47	AA	GG	-17.11433793
HS147110	LG14	27	Huile	VER09_NI	1.50E-06	0.06	52.32	AA	GG	1.507984568
HS152947	LG14	27	Huile	VER09_NI	8.84E-06	0.06	52.32	CC	TT	-1.413324867
HS152949	LG14	27	Huile	VER09_NI	8.84E-06	0.06	52.32	CC	TT	-1.413324867
HS057291	LG14	31.1	Huile	VE09_I	1.17E-05	0.06	53.23	AA	GG	1.344547918
HS061187	LG14	31.8	M0	SE10	1.30E-05	0.06	216.42	GG	TT	-1.992891454
HS070163	LG14	33.1	PMG	CO08_NI	1.62E-05	0.06	43.70	CC	TT	-1.260419752
HS155451	LG14	33.7	IF40	VER09_NI	5.70E-05	0.05	4.03	CC	TT	0.283001414
HS114137	LG14	34.6	PMG	CO08_NI	4.38E-06	0.06	43.70	CC	TT	-1.28468385
HS123988	LG14	34.7	PMG	CO08_NI	2.29E-05	0.05	43.70	GG	TT	1.247431518
HS148304	LG14	35	H2O	VE09_I	0.00031295	0.04	6.10	CC	TT	0.714844678
HS148306	LG14	35	H2O	VE09_I	0.00031295	0.04	6.10	AA	GG	0.714844678
HS069788	LG14	39.5	Hauteur	AI08_I	5.49E-06	0.06	148.60	AA	CC	7.123257962
HS138700	LG14	39.5	Hauteur	AI08_I	3.78E-05	0.05	148.60	CC	TT	6.843127243
HS115963	LG14	39.5	Hauteur	AI08_I	3.78E-05	0.05	148.60	CC	TT	6.843127243
HS118122	LG14	44.2	M3	CO08_I	1.02E-05	0.06	254.38	AA	CC	1.542018619
HS059912	LG15	8.7	H2O	AI09_NI	0.00030952	0.04	6.05	AA	TT	1.346104163
HS059912	LG15	8.7	M3	CO08_I	5.58E-05	0.05	254.38	AA	TT	1.923360867
HS059912	LG15	8.7	H2O	VE09_I	0.00023467	0.04	6.10	AA	TT	1.092790131
HS099019	LG15	9.7	H2O	AI09_NI	9.41E-05	0.04	6.05	AA	TT	1.192067886
HS148291	LG15	9.7	M3	CO08_I	3.90E-05	0.05	254.38	AA	GG	1.79818057
HS099019	LG15	9.7	H2O	CO08_I	1.41E-06	0.07	6.57	AA	TT	0.90294725
HS148291	LG15	9.7	H2O	CO08_I	3.97E-05	0.05	6.57	AA	GG	0.876882827
HS099019	LG15	9.7	M3	CO08_NI	6.05E-05	0.05	252.54	AA	TT	2.022910219
HS148291	LG15	9.7	M3	CO08_NI	5.92E-05	0.05	252.54	AA	GG	2.271771531
HS059538	LG15	9.7	H2O	GA09_NI	2.97E-05	0.05	8.61	GG	TT	-0.96353915
HS148291	LG15	9.7	H2O	VE09_I	0.00033765	0.04	6.10	AA	GG	0.941462361
HS101697	LG15	17.1	H2O	AI09_NI	8.75E-07	0.07	6.05	AA	CC	-1.730739344
HS122361	LG15	17.1	H2O	AI09_NI	2.14E-08	0.09	6.05	AA	GG	-1.992789017
HS085663	LG15	17.1	H2O	AI09_NI	6.75E-06	0.06	6.05	CC	TT	-1.929580041
HS101697	LG15	17.1	M3	CO08_I	7.52E-06	0.06	254.38	AA	CC	-1.925464794
HS149207	LG15	17.1	H2O	CO08_I	9.36E-06	0.06	6.57	AA	CC	-0.684040146
HS101697	LG15	17.1	M3	CO08_NI	1.00E-06	0.07	252.54	AA	CC	-2.716999312
HS122361	LG15	17.1	M3	CO08_NI	2.28E-05	0.05	252.54	AA	GG	-2.394506258
HS085663	LG15	17.1	M3	CO08_NI	3.83E-05	0.05	252.54	CC	TT	-2.79040092
HS101697	LG15	17.1	H2O	CO08_NI	1.63E-05	0.05	13.45	AA	CC	-1.682565307
HS101697	LG15	17.1	H2O	VE09_I	3.29E-06	0.06	6.10	AA	CC	-1.285135774
HS122361	LG15	17.1	H2O	VE09_I	5.12E-05	0.05	6.10	AA	GG	-1.137285669
HS085663	LG15	17.1	H2O	VE09_I	2.40E-05	0.05	6.10	CC	TT	-1.414180301
HS122361	LG15	17.1	RDT	VE10	1.75E-05	0.06	34.40	AA	GG	-2.332820037
HS096891	LG15	22	LAD	CO09_NI	8.56E-05	0.05	40.36	CC	TT	1.383544217
HS149260	LG15	22	LAD	CO09_NI	1.60E-05	0.06	40.36	AA	GG	-1.46548234
HS057257	LG15	40	H2O	AI08_I	2.31E-05	0.05	7.54	AA	GG	1.704260579
HS057257	LG15	40	H2O	AI08_NI	3.61E-06	0.06	9.09	AA	GG	2.195678976
HS057257	LG15	40	M3	CO08_NI	0.00016077	0.04	252.54	AA	GG	2.582539565
HS124776	LG15	47	H2O	VE09_I	0.00012956	0.04	6.10	AA	CC	-1.104984417
HS100931	LG15	47	IF20	VER09_NI	1.96E-06	0.07	6.12	GG	TT	-0.428329409
HS100931	LG15	47	Senescevol	VER09_NI	0.00014465	0.04	155.56	GG	TT	-13.14717279
HS087097	LG15	48.1	H2O	AI09_NI	0.00024375	0.04	6.05	CC	TT	-0.874011604
HS153812	LG15	48.1	H2O	AI09_NI	1.83E-05	0.05	6.05	AA	GG	-0.97728819
HS153812	LG15	48.1	H2O	VE09_I	6.49E-05	0.05	6.10	AA	GG	-0.700435243
HS087239	LG15	48.1	M3	VE10	1.34E-05	0.06	256.46	AA	GG	-2.048692807
HS104444	LG15	73	IF40	CA10	7.39E-05	0.05	0.51	CC	TT	-0.3473753
HS063124	LG16	53.5	H2O	AI09_I	0.00012891	0.04	6.86	AA	GG	-1.125672455
HS150602	LG16	77.7	H2O	CO08_I	5.90E-05	0.05	6.57	CC	TT	0.854891039
HS062236	LG16	84.2	F1	CO08_I	5.01E-05	0.05	212.89	AA	TT	1.173069846
HS149264	LG16	87.3	H2O	CO09_I	5.35E-05	0.05	6.34	CC	TT	0.358856593
HS083799	LG17	34	H2O	AI09_NI	0.00013112	0.04	6.05	CC	TT	-0.741907521
HS083799	LG17	34	M3	CO08_NI	3.33E-05	0.05	252.54	CC	TT	-1.301441449
HS083799	LG17	34	H2O	GA09_NI	3.40E-05	0.05	8.61	CC	TT	-0.716747945



HS083799	LG17	34	H2O	VE09_I	9.37E-06	0.06	6.10	CC	TT	-0.685346489
HS083799	LG17	34	IF40	CA10	2.18E-05	0.05	0.51	CC	TT	-0.365133258
HS134456	LG17	49.3	IF20	CO08_NI	2.50E-05	0.05	4.94	AA	TT	0.092260953
HS075890	LG17	49.3	IF20	CO08_NI	6.03E-06	0.06	4.94	AA	CC	0.091370493
HS143857	LG17	49.3	IF20	CO08_NI	1.29E-05	0.06	4.94	AA	CC	-0.094748839
HS150404	LG17	49.3	IF20	CO08_NI	0.00015581	0.04	4.94	AA	GG	-0.077940317
HS150977	LG17	49.3	IF20	CO08_NI	6.33E-05	0.05	4.94	CC	TT	0.0857184
HS150990	LG17	49.3	IF20	CO08_NI	6.33E-05	0.05	4.94	GG	TT	0.0857184
HS160458	LG17	49.3	IF20	CO08_NI	0.0001165	0.05	4.94	CC	TT	0.074776681
HS136835	LG17	52.7	H2O	AI09_I	6.20E-05	0.05	6.86	GG	TT	-1.224971142
HS136835	LG17	52.7	H2O	AI09_NI	0.00012516	0.04	6.05	GG	TT	-1.026954653
HS147573	LG17	53.3	H2O	AI09_NI	3.39E-06	0.06	6.05	AA	GG	-1.215131046
HS150073	LG17	53.3	H2O	AI09_NI	3.39E-06	0.06	6.05	CC	TT	1.215131046
HS147573	LG17	53.3	H2O	VE09_I	9.33E-06	0.06	6.10	AA	GG	-0.951825364
HS150073	LG17	53.3	H2O	VE09_I	9.33E-06	0.06	6.10	CC	TT	0.951825364
HS147559	LG17	53.3	F1	CA10	0.00012908	0.04	196.47	CC	TT	1.961710282
HS147564	LG17	53.3	F1	CA10	0.00012908	0.04	196.47	GG	TT	1.961710282
HS147573	LG17	53.3	F1	CA10	4.03E-05	0.05	196.47	AA	GG	-1.854468265
HS150073	LG17	53.3	F1	CA10	4.03E-05	0.05	196.47	CC	TT	1.854468265
HS150074	LG17	53.3	F1	CA10	0.00012908	0.04	196.47	CC	TT	1.961710282
HS150079	LG17	53.3	F1	CA10	0.00012908	0.04	196.47	GG	TT	1.961710282
HS147559	LG17	53.3	H2O	CHA10	1.80E-05	0.05	10.52	CC	TT	1.606101899
HS147564	LG17	53.3	H2O	CHA10	1.80E-05	0.05	10.52	GG	TT	1.606101899
HS147573	LG17	53.3	H2O	CHA10	5.96E-08	0.08	10.52	AA	GG	-1.919878024
HS150073	LG17	53.3	H2O	CHA10	5.96E-08	0.08	10.52	CC	TT	1.919878024
HS150074	LG17	53.3	H2O	CHA10	1.80E-05	0.05	10.52	CC	TT	1.606101899
HS150079	LG17	53.3	H2O	CHA10	1.80E-05	0.05	10.52	GG	TT	1.606101899
HS147573	LG17	53.3	H2O	SE10	5.76E-07	0.07	6.38	AA	GG	-0.44528662
HS150073	LG17	53.3	H2O	SE10	5.76E-07	0.07	6.38	CC	TT	0.44528662
HS111558	LG17	56.3	H2O	AI09_NI	0.00015572	0.04	6.05	CC	TT	0.980138557
HS160167	LG17	56.3	H2O	AI09_NI	0.00029181	0.04	6.05	CC	TT	-0.919433975
HS088134	LG17	56.3	H2O	VE09_I	0.00024	0.04	6.10	CC	TT	0.881949288
HS111558	LG17	56.3	F1	CA10	2.53E-06	0.06	196.47	CC	TT	2.118000263
HS111558	LG17	56.3	IF40	CA10	3.16E-05	0.06	0.51	CC	TT	0.405669253
HS088134	LG17	56.3	RDT	CHA10	1.42E-05	0.06	25.75	CC	TT	1.405424847
HS111558	LG17	56.3	RDT	CHA10	2.33E-05	0.06	25.75	CC	TT	1.23521608
HS160167	LG17	56.3	RDT	CHA10	8.60E-05	0.05	25.75	CC	TT	-1.118819127
HS111558	LG17	56.3	M3	SE10	7.62E-06	0.06	247.65	CC	TT	2.366954906
HS111558	LG17	56.3	H2O	SE10	1.59E-07	0.08	6.38	CC	TT	0.403358023
HS160167	LG17	56.3	H2O	SE10	1.45E-05	0.05	6.38	CC	TT	-0.337825787
HS137336	LG17	60	H2O	AI09_I	1.22E-05	0.06	6.86	AA	GG	-1.422365796
HS060553	LG17	60	H2O	AI09_I	2.25E-06	0.07	6.86	AA	CC	-1.59055352
HS137336	LG17	60	H2O	AI09_NI	2.79E-05	0.05	6.05	AA	GG	-1.187641142
HS060553	LG17	60	H2O	AI09_NI	6.91E-06	0.06	6.05	AA	CC	-1.318169882
HS137336	LG17	60	H2O	VE09_I	6.71E-05	0.05	6.10	AA	GG	-0.903035956
HS060553	LG17	60	H2O	VE09_I	2.15E-05	0.05	6.10	AA	CC	-0.981571703
HS137336	LG17	60	F1	CA10	1.31E-05	0.06	196.47	AA	GG	-2.255502785
HS060553	LG17	60	F1	CA10	5.35E-06	0.06	196.47	AA	CC	-2.362473406
HS137336	LG17	60	IF40	CA10	5.00E-05	0.06	0.51	AA	GG	-0.455737823
HS060553	LG17	60	IF40	CA10	2.13E-05	0.06	0.51	AA	CC	-0.475074625
HS137336	LG17	60	RDT	CHA10	3.58E-05	0.06	25.75	AA	GG	-1.395584949
HS060553	LG17	60	RDT	CHA10	1.07E-05	0.06	25.75	AA	CC	-1.475889478
HS137336	LG17	60	H2O	SE10	4.35E-05	0.05	6.38	AA	GG	-0.36671754
HS060553	LG17	60	H2O	SE10	3.08E-05	0.05	6.38	AA	CC	-0.371784565
HS053393	LG17	60.1	IF40	AI08_NI	7.97E-06	0.06	2.48	GG	TT	-0.493735599
HS062427	LG17	60.1	RDT	CHA10	2.54E-05	0.06	25.75	AA	TT	1.364750756
HS062427	LG17	60.1	H2O	SE10	6.65E-07	0.07	6.38	AA	TT	0.43959247
HS127609	LG17	62.1	LAI_F1	CA10	5.70E-05	0.05	4.42	CC	TT	0.091195062
HS158032	LG17	78.2	IF30	VE09_I	3.80E-07	0.08	6.15	AA	GG	0.358135669
HS158032	LG17	78.2	Senescevol	VE09_I	3.56E-05	0.05	167.47	AA	GG	12.1009156
HS164923	LG17	79.8	Huile	GA09_NI	1.36E-05	0.06	45.88	AA	GG	1.362137257
HS165182	x	x	F1	AI09_NI	3.99E-05	0.05	189.09	CC	TT	-1.706839133
HS131353	x	x	M3	CO08_I	3.17E-05	0.05	254.38	AA	GG	1.414969405
HS053096	x	x	H2O	CO08_I	0.00011658	0.04	6.57	CC	TT	0.978871832
HS052062	x	x	LAD	CO09_NI	4.15E-05	0.05	40.36	AA	CC	2.265771167
HS052506	x	x	LAD	CO09_NI	2.18E-06	0.07	40.36	AA	GG	2.025053321
HS053096	x	x	H2O	VER09_NI	8.13E-06	0.06	5.62	CC	TT	1.520251422
HS147860	x	x	H2O	SE10	2.18E-08	0.09	6.38	AA	TT	0.662568575

HS147876	x	x	H2O	SE10	8.58E-07	0.07	6.38	AA	GG	0.509111143
HS147877	x	x	H2O	SE10	7.67E-06	0.06	6.38	CC	TT	0.457431159
HS147878	x	x	H2O	SE10	7.67E-06	0.06	6.38	CC	TT	-0.457431159

**Table S0 : Statistiques d'association pour les différentes combinaisons (traits + environnements)**