Decoding the complexity of natural variation for shoot growth and response to the environment in Arabidopsis thaliana

Charlotte Trontin

UNIVERSITÉ PARIS-SUD

ÉCOLE DOCTORALE SCIENCES DU VÉGÉTAL

Laboratoire INSTITUT JEAN PIERRE BOURGIN

Discipline BIOLOGIE

THÈSE DE DOCTORAT

Soutenue le 21 Mai 2013

Par

Charlotte TRONTIN

Decoding the complexity of natural variation for shoot growth and response to the environment in *Arabidopsis thaliana*

Composition du jury :

Directeur de thèse	Olivier Loudet	DR2, INRA Versailles
Rapporteurs	Marie-Anne Félix	PR, ENS de Paris
	Frédéric Revers	CR1, INRA Bordeaux
Examinateurs	Maud Tenaillon	CR1, CNRS Ferme du Moulon
	Yvon Jaillais	CR1, CNRS ENS de Lyon
Présidente du jury	Jacqui Shykoff	DR1, CNRS Université Paris-Sud

CONTENTS

Table of contents 5

Acknowledgments 7

PhD thesis overview in French 9

Part I General Introduction 17

1. Arabidopsis thaliana as a model to study natural variation . . . 19
1.1 *A. thaliana* population biology . . . 19
1.1.1 *A. thaliana* taxonomy . . . 19
1.1.2 *A. thaliana* biogeography . . . 20
1.1.3 *A. thaliana* accessions . . . 21
1.1.4 *A. thaliana* life cycle . . . 22
1.2 Genetic & epigenetic diversity . . . 23
1.2.1 *A. thaliana* genetic variation . . . 23
1.2.1.1 Structure and demography . . . 24
1.2.1.2 The pattern of genetic variation . . . 26
1.2.1.3 Large scale genetic variation . . . 29
1.2.2 *A. thaliana* epigenetic variation . . . 30
1.3 Phenotypic diversity . . . 34
1.3.1 Quantitative phenotypic variation . . . 34
1.3.2 Phenotyping quantitative traits . . . 35
1.3.3 Integrated phenotypic variation and plasticity . . . 37

2. From phenotypic variation to molecular variants and vice versa . . . 41
2.1 Mapping and cloning the genes responsible for phenotypic variation . . . 41
2.1.1 Linkage mapping in segregating population . . . 42
2.1.1.1 Mapping populations . . . 42
2.1.1.2 Statistical methods . . . 45

2.1.1.3 Conclusion on the use of linkage mapping . . . 46
2.1.2 Cloning strategies . . . 47
2.1.3 Genome-Wide Association mapping . . . 49
2.1.4 Nested association mapping (NAM) . . . 51
2.1.5 Validation of causal genes and polymorphisms . . . 52
2.2 The complex genetic architecture of quantitative traits . . . 54
2.2.1 The number and effects of QTLs . . . 54
2.2.2 Rare variants & allelic heterogeneity . . . 56
2.2.3 Pleiotropy or linkage . . . 57
2.2.4 Epistatic interactions . . . 58
2.3 The unexpected complexity of phenotypic variation. . . . 59
2.3.1 Phenotypic buffering . . . 59
2.3.2 The genetic x environment interactions . . . 60
2.4 Conclusion about QTLs detection using natural variation. . . . 61

3. Evolutionary significance of A. thaliana natural variation . . . 63
3.1 Ecologically significant phenotypic variation in *A.thaliana.* . . . 63
3.2 Evolutionary significance of phenotypic variation . . . 69
3.2.1 Theoretical elements . . . 69
3.2.2 Methods to detect evolutionary significant variants . . . 72
3.2.2.1 Tests based on within-species variation . . . 74
3.2.2.2 Tests based on between-species variation . . . 77
3.3 Conclusion: Major evolutionary traits in *A. thaliana*? . . . 78

4. Scope of the thesis . . . 81

Part II Allelic heterogeneity and trade-off shape natural variation for response to soil micronutrient 83

5. Project background and personal contribution . . . 85

6. Publication in PLoS Genetics . . . 89

7. General discussion and Perspectives . . . 109
7.1 Why is *MOT1[Sha]* hypofunctional? . . . 109
7.2 Is *MOT1[Sha]* adaptive? . . . 111
7.3 What are the differences between MOT1[Sha] and MOT1[Ler] ? . . . 118
7.4 Further investigations . . . 119

Part III A tandem of receptor-like kinases is responsible for natural variation in shoot growth response to mannitol treatment in A. thaliana 123

8. Project background and personal contribution . . . 125

9. Publication . . . 127

10. General discussion and Perspectives . . . 153
10.1 *EGM1* and *EGM2* and biotic stress . . . 153
10.1.1 Innate immunity in plants . . . 153
10.1.2 Mannitol signalling pathway in plant immunity . . . 154
10.1.3 Which pathogens and in which plants? . . . 157
10.2 What are the roles of EGM receptor-like-kinases within mannitol signalling pathway? . . . 160
10.2.1 The role of plant receptor like kinases in plant innate immunity . . . 160
10.2.2 What is the function of EGMs RLKs? . . . 161
10.3 *EGM1* and *EGM2* natural variation and evolution. . . . 168

Part IV Natural epigenetic variation at QQS loci in A. thaliana 171

11. Project background and personal contribution . . . 173

12. Publication in PLoS Genetics . . . 175

13. General discussion and Perspectives . . . 189
13.1 The phenotypic consequences of *QQS* epivariants . . . 189
13.2 The origin and evolution of *QQS* epivariants . . . 195

Conclusion 203

Appendix 207

A. Additional publication . . . 209

Bibliography 211

ACKNOWLEDGMENTS

The work presented in that thesis is far from being the work of a single person and I would like to acknowledge all the people that contribute to it. Many thanks to:

Olivier Loudet for giving me the opportunity to work on all the projects I contributed to and for its general trust, support and advices.

Kian Poormohammad for all the hard work he did on MOT1 and EGM projects.

Matthieu Simon (*Institut Jean Pierre Bourgin*) & **Thierry Robert** (*Écologie, Systématique & Évolution*) for their help with *MOT1* population genetics analyses.

David Salt & **Matthew Andreatta** (*University Aberdeen and Purdue University*) for their expertise and for the discussions we had regarding MOT1 transporter, plant Mo biology and soil Mo contents.

Dan Kliebenstein & **Jason Corwin** (*Davis University*) for performing biotic stress experiments and for commenting the EGM manuscript.

Dominique Roby & **Claudine Balagué** (*Laboratoire des Intéractions Plantes-Microorganismes*) for performing biotic stress experiments and providing various hormonal mutants.

Michel Vincentz, **Amanda Silveira** (*Center for Molecular Biology and Genetic Engineering*) & **Vincent Colot** (*Institut de Biologie de l'École Normale Supérieure*) for involving me and Olivier in the QQS project.

Kian Hematy, **Nathalie Vrielynck**, **Halima Morin**, **Lionel Gissot** & **Magali Bedu** (*IJPB*) for specific technical assistance and advices.

Lilian Dahuron, **Jean Sébastien Ferré** & **Bernard Josselin** (*IJPB*) for taking care of my plants in the greenhouse.

I am also really grateful to all my labmates for their support, technical assistance, discussions, advice and for all the great times I had in their company. More particularly, thanks to **Marina**, **Elo** & **Lien** for their shoulders, smiles and fun (et pleins d'autres trucs trop long à énumérer), to '**Petit Matthieu**' for sharing is bench and pipets with me and for its technical assistance, to **Évelyne**, **Carine**, **Vincent** & **Cécile** for the relaxing tea breaks (et Carine et Laurent pour les délicieux macarons), to '**M4**' for taking care of my house & plants while I was away and for its support with **Elo2** & **Yann** during the last months of writing, to **Francisco** & **Sébastien** for being so fond of my 'tiramisu' and 'tarte aux pralines' (enfin ça c'est aussi vrai pour tous les autres), to **Christine** for her interest regarding ikebana (and for the wonderful book), to **Nico** & **Isa** for fun coffee times, to **Alex** for its support as a PhD student and for the great moments we had in the various conferences (la prochaine fois on réussira à leur soutirer plus d'infos à ces chinois ;-)) and to all the current and former members of the lab, more particularly **Dana** & **Val**.

Finally, I would like to thank my family and Sebastien for their support and love.

During my PhD, I was funded by the École Normale Supérieure de Lyon and by a studentship from the French Ministry of Research.

PHD THESIS OVERVIEW IN FRENCH

Décoder la complexité de la variabilité naturelle pour la croissance et la réponse à l'environnement chez *Arabidopsis thaliana*

Contexte scientifique

Des génotypes adaptés à des environnements contrastés ont de grandes chances de se comporter différemment lorsqu'ils sont placés dans des conditions similaires et contrôlées, notamment si leur sensibilité aux signaux environnementaux et/ou leur croissance intrinsèque sont limitées à différents niveaux. De ce fait, la variabilité observée dans les populations naturelles peut être utilisée comme une source illimitée de nouveaux allèles ou gènes pour l'étude des bases génétiques de la variation des traits quantitatifs. Le but des approches de génétique quantitative est de comprendre comment la diversité génétique et épigénétique contrôle la variabilité phénotypique observée dans les populations à différentes échelles, au cours du développement et sous différentes contraintes environnementales. De plus, ces analyses ont pour objectif de comprendre comment les processus adaptatifs et démographiques influencent la fréquence de ces variants dans les populations en fonction de leur environnement local.

Arabidopsis thaliana est une des plantes modèles les plus étudiées en raison de son cycle de vie court, de sa petite taille, de sa production importante de graines et de l'efficacité de la transformation de cette espèce. De plus la séquence complète de son petit génome est disponible ainsi que de nombreuses lignées mutantes. En ce qui concerne la variabilité naturelle, c'est également un très bon modèle puisque cette petite herbacée est retrouvée dans des habitats relativement contrastés, dans le monde entier. Son potentiel adaptatif particulièrement élevé est aussi suggéré par l'importante diversité génétique, épigénétique et phénotypique observée chez cette espèce. Le terme d' 'accession' est utilisé pour faire référence à des individus collectés sur un site donné.

L'étude de la variabilité naturelle peut être appréhendée en utilisant diverses approches. Tout d'abord, afin d'identifier les loci (on parle de QTL pour *Quantitative Trait Loci*), les gènes (QTG) ou même les polymorphismes (QTP) responsable de la variabilité phénotypique complexe observée entre différentes accessions dans des conditions environnementales données, des

approches de génétique quantitative sont utilisées. Ces approches sont basées sur l'association entre des données génétiques et des données phénotypiques issues du génotypage et du phénotypage de populations de cartographies telles que des F2 ou des RILs ou directement des accessions. Dans le premier cas on parle en général de cartographie de QTL (*QTL mapping*) et dans le deuxième cas on parle de génétique d'association (*Genome-Wide Association mapping ou GWA*). L'identification des bases génétiques à l'origine des variations des traits quantitatifs observées dans les populations naturelles peut être plus ou moins difficile en fonction de l'architecture génétique des traits en question et de leur intéraction avec l'environnement. Suite à l'identification et la confirmation des QTGs, il est possible d'utiliser des approches plus classiques de biologie moléculaire et cellulaire et de physiologie pour comprendre la fonction du QTG en relation avec la variation phénotypique observée et l'impact du/des polymorphisme(s) en cause sur la fonction du gène. Par ailleurs, l'identification des QTGs et QTPs permet de rechercher des traces sélections au niveau du loci d'intérêt en utilisant des approches plus évolutives. Enfin des corrélations entre la fréquence des variants génétiques, leurs effets sur la fitness et les facteurs environnementaux charactérisant les régions dans lesquelles on retrouve ces variants peuvent aussi être très utiles pour comprendre comment les populations s'adaptent localement.

Mon travail de doctorat a consisté en l'analyse de la variabilité naturelle pour la croissance et la réponse à l'environnement chez *A. thaliana*. J'ai eu la chance de participer à trois projets indépendants qui exploitent tous la variabilité naturelle d'*A. thaliana* et qui m'ont permis d'aborder les différentes approches disponibles pour l'étude des variations des traits quantitatifs dans les populations naturelles.

Validation de l'évolution non-neutre du gène *MOT1* codant pour un transporteur de molybdate

Ce projet a été initié par Olivier Loudet en 1999 suite à l'observation que l'accession Shahdara (ou Sha) (collectée dans la Vallée Shakhdara au Tajikistan) montrait une réduction de croissance bien plus important que l'accession Bay-0 (collectée à Bayreuth en Allemagne) quand ces accessions étaient cultivées sur un milieu riche en tourbe et donc relativement acide. En utilisant une population de RILs et des lignées de HIFs, Claire Le Metté et Kian Poormohammad ont montré qu'un locus majeur de 80kb, en haut du chromosome 2 expliquait cette différence de croissance entre les deux accessions. Cet interval génétique contenait parmi 18 autres, le gène *MOT1* codant pour un transporteur de molybdate (la forme du molybdène (Mo) assimilable par la plante). Ce gène était un très bon candidat pour le QTL de croissance observé sur tourbe car la disponibilité en Mo pour la plante varie en fonction du pH du sol et une

déficience en Mo, cofacteur essentiel à de nombreuses enzymes dont la nitrate reductase, peut conduire à des phenotypes similaires à ceux que présente l'accession Sha sur tourbe. Au cours de son post-doc, Kian a montré qu'une mutation dans l'allèle Sha du gène *MOT1* (D104Y) rend ce transporteur hypofonctionnel et que le phénotype 'tourbe' est bien le résultat d'une carence en Mo. Dans deux études de 2007 et de 2008, une délétion de 53 paires de bases (Del-53-pb) dans le promoteur du gène *MOT1* a été observée chez plusieurs accessions (dont Ler-0 et Van-0) et a été associée à une réduction du niveau d'expression du gène *MOT1* ainsi qu' à une diminution du contenu en Mo dans les feuilles et les racines de ces accessions. Partant de ce constat, Kian a montré que toutes les accessions de type 'Ler' (Del-53-pb) présentaient aussi le phénotype 'tourbe' et les données de contenu en Mo disponibles ont par ailleurs montré que toutes les accessions de type 'Sha' (D104Y) présentent un contenu en Mo réduit en condition standard. Ainsi au moins deux allèles différents du gène *MOT1* ségrègent dans les populations naturelles d'*A. thaliana* et sont responsables d'un phénotype similaire à savoir une réduction de la quantité de Mo dans la plante ainsi qu'une réduction de croissance sur des sols pauvres en Mo. Cette forme d'hétérogénéité allélique est probablement assez fréquente bien que peu décrite.

Quand je suis arrivée dans l'équipe, l'importance évolutive de l'hétérogénéité allélique observée à *MOT1* était mal connue. Au cours de ma thèse, j'ai réalisé avec l'aide de Matthieu Simon et de Thierry Robert des analyses de génétique des populations dans le but de détecter de possibles traces de sélection au niveau du locus *MOT1*. Le séquencage de 102 accessions d'*A. thaliana* réparties dans tout l'hémisphère nord, dont 28 Sha-like, 19 Ler-like et 55 Col-like a montré que l'allèle Sha présente un faible niveau de polymorphismes comparé à *MOT1[Ler]* et *MOT1[Col]*. Ce faible taux de polymorphismes a aussi été observé à une échelle régionale en 'Asie de l'ouest' où sont confinées les accessions de type Sha et peut refléter une expansion récente et rapide de l'allèle Sha dans cette région. Cette expansion peut être le résultat de processus évolutifs neutres au cours de la recolonisation postglaciaire de 'Asie de l'ouest' tel que du *'gene surfing'* ou peut refléter des évènements de sélection locale en faveur de l'allèle Sha. Cette dernière hypothèse est en accord avec la significativité de plusieurs tests de sélection réalisés sur un échantillon mondial et régional d'accessions, dont le test de Tajima ($D<0$), le test de Mc Donald-Kreitman ($NI>1$) et le test HKA, suggérant que la diversité génétique observée à *MOT1* est le résultat de la sélection de différents haplotypes. De plus, l'analyse élémentaire de plusieurs sols sur lesquels des accessions d''Asie de l'ouest' ont été collectées a montré que les accessions de type Sha poussaient sur des sols relativement riches en Mo. Enfin, des analyses de toxicité effectuées sur *A. thaliana* en serre et *in vitro* ont montré que de fortes concentrations en Mo diminuaient la croissance racinaire et le nombre de graines produites par plante. Ces tests ont aussi montré que les plantes portant un allèle défectueux à *MOT1* produisaient des

graines en moyenne légèrement plus grosses que les plantes de type sauvage et ce spécifiquement sur des sols très riches en Mo. Ainsi, l'allèle Sha pourrait avoir à la fois un effet protecteur en réponse à de fortes accumulations de Mo dans l'environnement et être délétère pour la plante sur des sols pauvres en Mo.

La découverte de variants génétiques associés à des phénomènes de trade-off environnementaux reste assez rare malgré leur importance dans l'évolution et la répartition des haplotypes au sein des populations et des espèces. De plus notre travail met en évidence l'importance des paramètres environnementaux hétérogènes, tel que les propriétés des sols, comme agents de sélection à l'origine de l'adaptation. Ce travail a été publié dans la revue PLoS Genetics le 12 Juillet 2012.

Deux récepteurs kinases dupliqués en tandem sont responsables de la variation de croissance observée en réponse à un traitement au mannitol dans des populations naturelles d'*A. thaliana*

Depuis ses débuts, le groupe VAST s'emploie à décoder les bases génétiques de la tolérance aux contraintes abiotiques dans des populations naturelles d'*A. thaliana* en utilisant des approches de génétique quantitative. Pour ce faire, des protocoles de phénotypage haut débit *in vitro* ont été développés. Le mannitol, un polyol connu pour ne pas être synthétisé par *A. thaliana* ni perturber son métabolisme, est généralement utilisé dans les milieux gélosés pour induire des stresses osmotiques. Dans cette étude, les bases génétiques de la variabilité pour la croissance en réponse au stress induit par un milieu contenant 60mM de mannitol ont été recherchées dans une population de RILs issus du croisement des accessions Col-0 et Cvi-0. Un QTL majeur et spécifique du milieu mannitol a été identifié en haut du chromosome 1 et sa ségrégation a été confirmée dans une famille de HIFs (pour '*Heterogeneous Inbred Family*'). Ce QTL a été nommé EGM pour '*Enhanced Growth under Mannitol stress*' (augmentation de croissance sous stress induit par le mannitol) à cause de l'effet positif de l'allèle Cvi du QTL sur la croissance en condition de stress induit par le mannitol. Kian Poormohammad, a cartographié à l'échelle du gène le QTL EGM et l'intervalle candidat de 10kb identifié a été validé en utilisant une famille de lignées appelées arHIFs pour '*advanced recombined HIF*' qui ségègent uniquement pour la région d'intérêt. L'analyse de plusieurs mutants T-DNA pour chacun des 3 gènes présents dans l'intervalle candidat a finalement suggéré qu'une mutation dans le gène *At1g11300* devait être la cause du QTL EGM (i.e le QTG).

En ce qui concerne ce projet, le but de ma thèse a été de confirmer le QTG, d'identifier le ou les polymorphismes responsable(s) du QTL et de comprendre la fonction du gène vis à vis du stress induit par le mannitol.

Tout d'abord, j'ai montré que le gène annoté *At1g11300* dans TAIR10 recouvre en réalité deux gènes paralogues indépendants apparus par duplication en tandem et codant pour des récepteurs kinases putatifs de la famille SD1 et caractérisés par un peptide signal, une région extracellulaire et un domaine kinase intracellulaire. Étant donné que les mutants T-DNA disponibles dans ces deux gènes sont plus grands que leur fond génétique sauvage en condition de stress induit par du mannitol, ces deux gènes ont été renomés *EGM1* (*At1g11300*) et *EGM2* (*At1g11305*).

Dans la région candidate, neuf polymorphismes différencient l'allèle Col de l'allèle Cvi, quatre d'entre eux étant non-synonymes. Afin d'identifier parmi ces SNPs celui responsable du QTL EGM, j'ai testé la ségrégation d'EGM dans plusieurs populations de F2 spécialement créées pour ségréger pour diverses combinaisons de polymorphismes de type Col et Cvi, une approche que l'on a appelé *génétique d'association spécifique*. Cette approche m'a permis de réduire le nombre de SNPs candidats à 3: une mutation de T→C dans le promoteur d'*EGM1*, une mutation non-synonyme d'une sérine conservée dans la famille SD1 en une glycine dans le domaine lectine d'*EGM2* et une mutation non-synonyme d'une cysteine très conservée et impliquée dans la formation d'un pont disulfure important pour le repliement du domaine PAN-APPLE d'*EGM2*. Des complémentations transgéniques ont montré que l'allèle Cvi du gène *EGM2* est au moins hypofonctionnel probablement du aux deux mutations non-synonymes indiquées ci-dessus. De plus des analyses d'expression des gènes EGMs dans les arHIFs, les mutants *egm1* et *egm2* et dans différentes accessions suggèrent que le polymorphisme dans le promoteur d'*EGM1* pourrait aussi contribuer au phénotype EGM. Finalement, certaines de nos données montrent que ces deux récepteurs kinase putatifs ne sont pas complètement redondants bien qu'impliqués dans un pathway commun. Le rôle exacte de chacun reste a déterminer.

Au début de notre étude, le mannitol a été utilisé pour induire un stress osmotique. Cependant, comme les deux gènes responsables du QTL présentent des domaines de liaison au mannose (domaine B-lectine), nous nous sommes demandés si le phénotype EGM que nous observions sur mannitol était bien le résultat d'un stress osmotique. Différentes contraintes osmotiques ont été testées mais la ségrégation d'EGM n'a été observée que sur mannitol suggérant que le mannitol en lui même pouvait avoir un effet sur la croissance indépendamment de la contrainte osmotique qu'il impose. Une analyse transcriptomique a montré qu'environ 200 gènes sont différentiellement exprimés entre les 2 arHIFs sur mannitol. Parmi eux, un enrichissement en gènes répondant à des contraintes biotiques a été observé, enrichissement qui a par ailleurs été montré dans une analyse comparant le transcriptome de plantes Col-0 traitées ou non au mannitol. De plus, l'induction de plusieurs de ces gènes de stresses biotiques est spécifique du milieu mannitol et n'est pas observé sur un milieu contenant du sorbitol. L'ensemble de ces observations nous ont amenés à penser que les gènes *EGM1* et *EGM2* pouvaient jouer un rôle dans la réponse aux contraintes biotiques. En effet, on sait que certains pathogènes, dont

les champignons, utilisent le mannitol comme source de carbone. De plus, quelques papiers suggèrent qu' au cours de l'infection, les pathogènes sécrètent du mannitol pour faire face à la production de ROS par la plante. En accord avec cette hypothèse, nous avons montré que les mutants *egm1* et *egm2* étaient plus sensibles à *Botrytis cinerea* que les fonds génétiques sauvages respectifs.

En conclusion, ce projet nous a permis d'identifier deux récepteurs kinase putatifs potentiellement impliqués dans les réponses de défense contre certains pathogènes via la perception directe ou indirecte du mannitol produit par ces derniers. Ce travail sera bientôt soumis dans une revue scientifique (PNAS ou Plant Cell).

La variabilité épigénétique observée au locus *QQS* dans des populations naturelles d'*A. thaliana*

Ce projet a été initié par Michel Vincentz (CBMEG, Brésil) et son étudiante en thèse Amanda Silveira. Alors qu'ils étudiaient un facteur de transcription potentiellement impliqué dans l'homéostasie cellulaire du carbone, ils ont observé par hasard que le gène *QQS* (pour *Qua-Quine Starch*) était spécifiquement et fortement up-régulé dans un de leur stock de graines Col-0, stock qu'ils ont appelé 'Col-0*'. En utilisant des approches de génétique et d'épigénétique, ils ont montré que différents épiallèles stables du gène *QQS* ségrègent dans leur stock de graines Col-0* et dans quelques accessions d'*A. thaliana*. Ces epiallèles sont caractérisés par une forte corrélation négative entre le niveau de méthylation du promoteur et du 5'UTR du gène *QQS* et le niveau d'accumulation de son transcrit. De plus ils ont montré que ces variations de méthylations sont indépendantes de la séquence génétique observée au niveau de *QQS* ainsi que de l'état de méthylation des transposons présents autour de ce gène. Partant de ces résultats, nous avons participé à trois aspects différents du projet.

Tout d'abord, plusieurs personnes au laboratoire travaillent sur l'identification et la confirmation des QTLs d'expression (eQTL) dans les populations de RILs issues des croisements Cvi-0 x Col-0 et Bur-0 x Col-0. Un des eQTLs locaux les plus significatifs identifié parmi ces deux populations colocalise avec le gène *QQS* dont il régule l'expression. Avec l'aide de Matthieu Canut et grâce à l'expertise du laboratoire dans ce domaine, nous avons confimé que *QQS* est bien controllé en cis par cet eQTL en utilisant des tests ASE (pour *Allele Specific Expression*) dans des F1 issus des croisements de Col-0 (qui présente un allèle méthylé) avec les accessions Jea et Kondara (qui présentent des allèles déméthylés comme Cvi-0). De plus nous avons montré qu'il était possible de restaurer l'expression de *QQS* dans des accessions présentant un allèle methylé en traitant les plantes avec un inhibiteur de méthylation. Ces

résultats ont fortement renforcé l'hypothèse selon laquelle les épivariants observés au gène *QQS* sont stables et 'pure' dans le sens ou ils ne semblent pas dépendre de variants génétiques.

Ensuite, à partir de graines collectées directement sur le terrain, en Asie Centrale, par Olivier Loudet, nous avons montré que les épiallèles au gène *QQS* ségrègent bien dans les populations naturelles d'*A. thaliana* et ne sont pas juste le résultat de plusieurs générations d'autofécondation en condition expérimentale en serre. Au sein de trois populations différentes, des variants méthylés et non méthylés ont été isolés posant la question du potentiel adaptatif de ces épivariants.

Enfin nous avons essayé de tester ce dernier en cherchant un phénotype associé aux différents épivariants (méthylé et non méthylé). Le gène *QQS* ayant précédemment été caractérisé comme un régulateur négatif de la synthèse d'amidon, nous avons testé deux phénotypes –la croissance et l'accumulation et dégradation de l'amidon au cours de 24h– dans différentes conditions –*in vitro* et *in vivo*, en condition normales ou en condition de contraintes hydrique, osmotique ou froid–. Cependant aucune différence significative n'a été observée entre les différentes lignées suggérant que les épivariants au gène *QQS* n'ont pas de conséquence phénotypique évidente (en tout cas dans nos conditions) et donc n'ont peut être pas de potentiel adaptatif.

En conclusion ce travail a permis la caractérisation précise d'un épiallèle 'pure' ségrégeant à une fréquence relativement élevée dans des populations naturelles d'*A. thaliana* –caractérisation qui chez les plantes reste rare–. Bien que l'importance phénotypique et écologique de ces épiallèles reste à démontrer, ce travail souligne la contribution des variations épigénétiques aux variations stables du niveau d'expression de certains gènes dans les populations. Enfin, ce travail suggère que les gènes nouvellement formés (neo-gènes) pourraient être particulièrement enclins aux variations de niveau de méthylation. Ce travail a été publié dans la revue PLoS Genetics le 11 Avril 2013.

Part I

GENERAL INTRODUCTION

1. *ARABIDOPSIS THALIANA* AS A MODEL TO STUDY NATURAL VARIATION

Arabidopsis thaliana or the mouse ear cress is one of the first plant model species. It has been widely used in genetics and molecular biology because of its short life cycle (about 6 weeks from germination to seed maturation in the best case), small size (so that it can be grown in restricted space) and high seed production. The genome sequence has been published in 2000 with extensive genetic and a nearly-perfect physical map. Besides, the numerous mutant lines available and the high transformation efficiency of this species strongly favoured its maintenance as an important model in plant biology.

Regarding natural variation, *A. thaliana* is also a very good model. First this weed is found in very diverse habitats all around the world and so likely has the capacity to quickly adapt to many local environmental constraints. Then, it presents a wide genetic and phenotypic diversity that might illustrate the huge plasticity of the species. In this section, I will independently review our current knowledge of A. thaliana genetic and phenotypic natural variation and how this diversity is constrained by the biology of the species and the environment (figure 1). I wrote a short review for Current Opinion in Plant Biology on similar grounds, which is placed in appendix of this manuscript.

1.1 A. thaliana population biology

In this section, I will give basic information regarding *A. thaliana* population biology that are necessary to understand *A. thaliana* genetic diversity.

1.1.1 A. thaliana taxonomy

Arabidopsis thaliana (L.) Heyne (2n=10) is a small annual flowering plant that belongs to the Brassicacea family which includes several important crop species such as mustard, cabbage, radish and rapeseed. It belongs to a small genus (Arabidopsis) including 8 other members [1] from which it diverged about 10Mya (figure 2). Beyond *A. thaliana*, three major lineages dividing into several subspecies, are generally recognized, namely *A. lyrata*, *A. halleri* and *A. arenosa* [1, 2, 3]. Compared to most of its close relatives, *A. thaliana* has the unique feature to be an essentially-selfing species. The selfing rate has been estimated around 97%

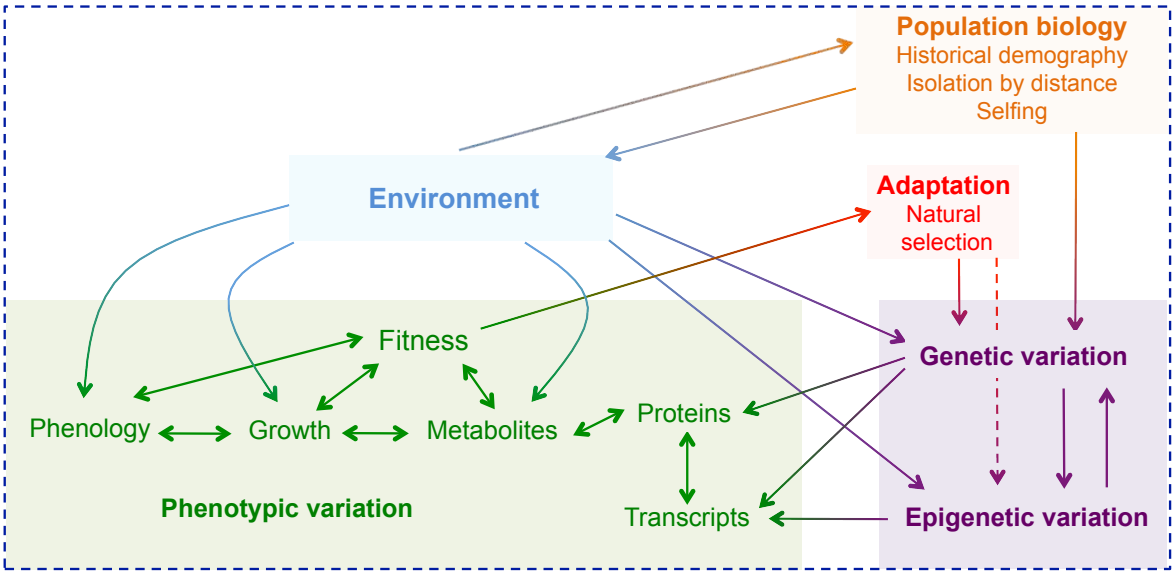


Fig. 1. Schematic recapitulation of the interplay between different factors influencing Arabidopsis natural variation and presented in this thesis. The different arrows represent effects of one factor on another. The arrow is dotted if the effect has not been shown in the literature.

but can vary locally in populations [4, 5]. Like other weedy inbreeding species, *A. thaliana* exhibits typical reduced unscented flowers with stigmas and anthers close together and a small genome with a reduced chromosome number [6]. The recent sequencing of the *A. lyrata* genome (2n=16; 200Mb) helped understanding the shrinkage of *A. thaliana* genome (2n=5, 150Mb) [7]. They showed that 10% of the genome size difference was due to large-scale rearrangements that caused the loss of 3 centromers in *A. thaliana*. But the 90% resting is attributed to hundreds of thousands small deletions. Those deletions are mostly found in non-coding DNA and transposons. This analysis raised the question of the selective advantage of a small genome and reduced chromosome number for a selfing species. It further illustrates that beyond the use of *A. thaliana* close relatives in the exploration of self-incompatibility (*A. lyrata*), heavy metal tolerance (*A. halleri*) and inter-specific hybridization (*A. arenosa*), those species have been and will be more and more useful in comparative genomics approaches that aim at understanding the evolution of genomes. Regarding *A. thaliana* natural variation, they are important to help define the ancestral state of the genetic variation observed within this species.

1.1.2 A. thaliana biogeography

A. thaliana is native from Eurasia (figure 3) where the climate is likely limiting the species distribution range: too low spring and autumn temperatures limit the distribution in Northern Europe and high temperatures with low precipitation in summer limit the distribution of the species in North Africa and in Southwest and Middle Asia [8]. Due to human-induced dispersal,

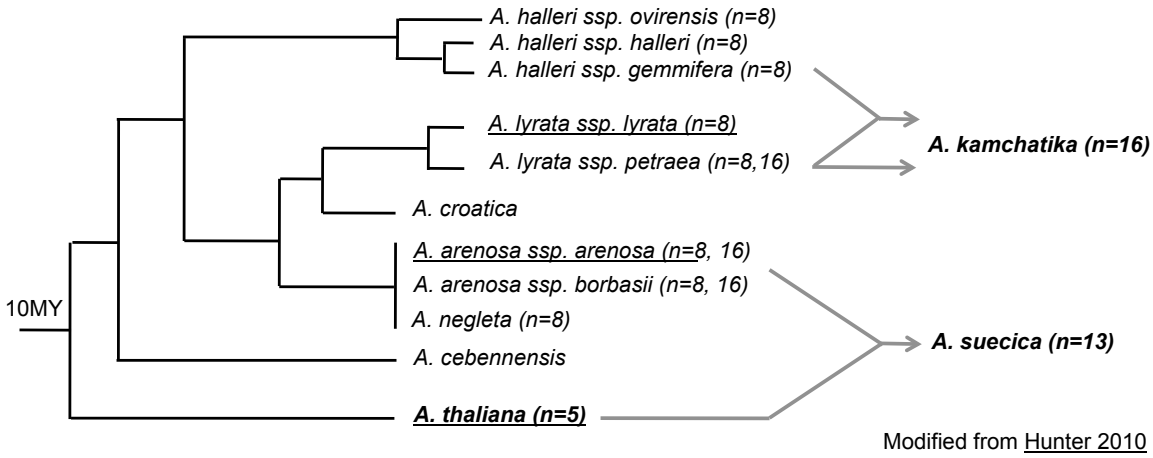


Fig. 2. Phylogenetic relationships within Arabidopsis genus based on Al-Shehbaz and O'Kane, 2002 and Hoffmann, 2005 [1, 9]. The tree was modified from Hunter 2010 [3]. Chromosome numbers from Al-Shehbaz and O'Kane, 2002 are given in parentheses. Grey arrows indicate parental species of allopolyploids. Note that *A. kamchatika* originated both as a hybrid of *A. halleri* x *A. lyrata*, and as a autopolyploidy derivate of *A. lyrata*. Selfing species are indicated in bold and species for which the genome was or is being sequenced are underlined. *A. thaliana* diverged from *A. lyrata* about 10MY ago [7] .

A. thaliana is now distributed in the whole northern hemisphere including north America. It can be found up to 4,250m in diverse open or disturbed habitats such as human-associated cultivated and waste areas, sparse meadows, rubble slopes, riverbanks, beaches and roadsides [1]. This range of distribution is much larger than the one of its close relatives, which may be reflecting differences in life cycle strategies. Nevertheless, if *A. thaliana* is able to quickly colonize various habitats, it is also a poor competitor in dense vegetation and so is also vulnerable to rapid extinction.

1.1.3 A. thaliana accessions

For the past 2 decades at least, thousands of *A. thaliana* individuals have been actively collected from diverse worldwide locations (figure 3). The term 'accession', rather than ecotype, is commonly used to refer to those individuals collected at a specific location, as it does not imply that the individual has a unique ecology nor that it is adapted to a specific environment [10]. Because of the selfing nature of *A. thaliana*, most of those individuals represent essentially-inbred lines, which are practically homozygous and can be maintained for several generations (but see paragraph on the pattern of genetic variation regarding lab-induced mutations). As a consequence, bulked or single-seed descendants of original lines are distributed by several

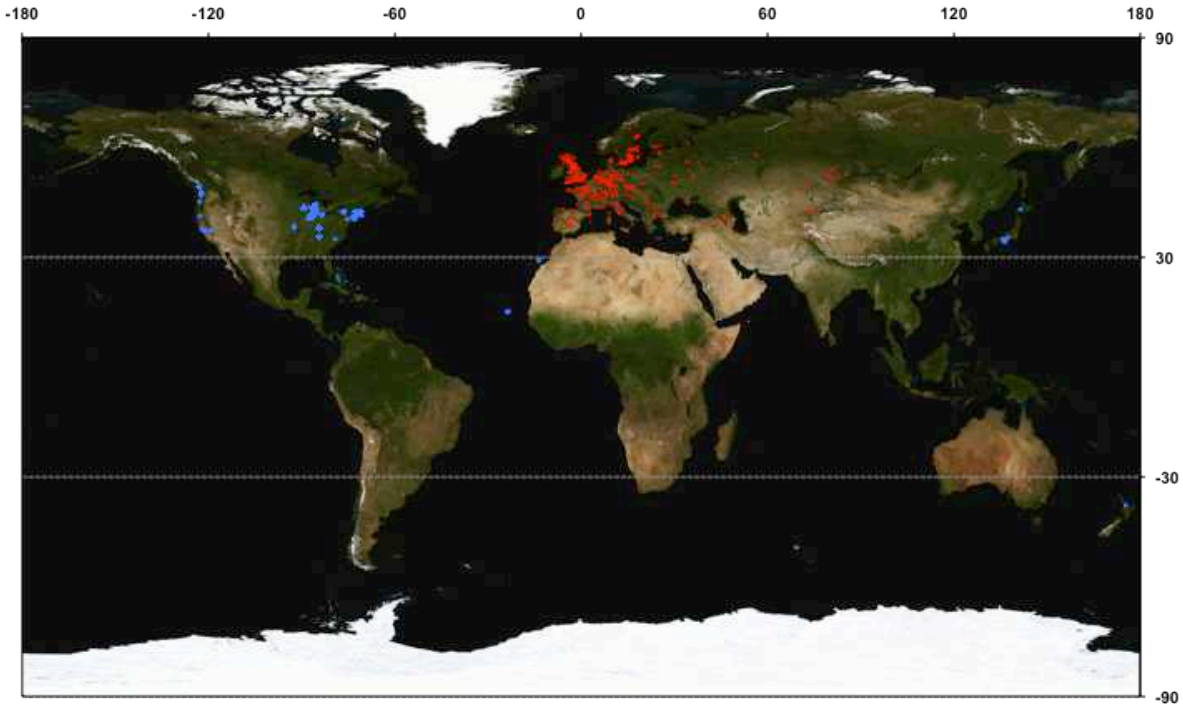


Fig. 3. Worldwide repartition of over 7000 collected *A. thaliana* accessions. Accessions located in the presume native range are indicated in red whereas introduction are in blue.

resource centers [1] to many different labs over the past decades. Besides, as fixed material, accessions need to be genotyped only once but can be phenotyped repeatedly for different traits and under various conditions.

1.1.4 A. thaliana life cycle

For plants, the timing of the major steps of the life cycle such as germination and flowering is crucial to ensure plant survival and reproduction. This timing is genetically determined but also depends on the environmental cues the plants are facing (figure 4). In Europe, *A. thaliana* flowers early in spring probably to minimize inter-specific competition. It produces seeds from April to July, sometimes later in the season until autumn. Regarding germination, two different life cycle strategies have been adopted by *A. thaliana* depending on the environment the plant is facing [10]. Some of the Southern and Western European accessions are summer annuals. They germinate in spring and complete their life cycle during the summer to spend winter as seeds. Others as well as most of Northern European accessions are winter annuals. They germinate in late summer or autumn, spend winter as rosettes and flower early in spring. Interestingly, within populations (i.e. at a specific location) mixes of winter and summer annuals can be

1. Arabidopsis Biological Resource Center (ABRC), the Nottingham Arabidopsis Stock Center (NASC), the RIKEN BioResource Center (RIKEN) and the Centre de Ressource Biologique (VASC)

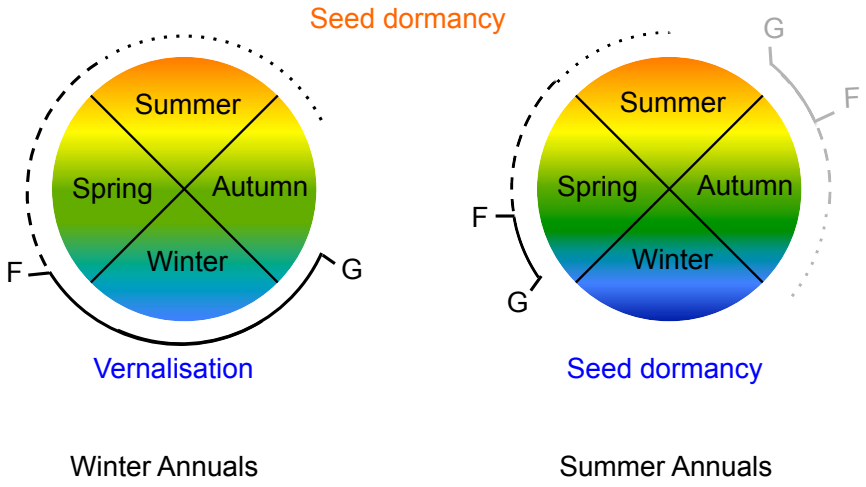


Fig. 4. *A. thaliana* life cycles. Winter annual accessions germinate (G) at fall, over-winter as rosettes and flower (F) early in spring. Summer annual accessions are cycling much more rapidly as they germinate (G) and flower (F) in spring. Depending on summer weather, those latter accessions could undergo a second life cycle during the year (in grey). Vernalisation and seed dormancy are important processes partly regulating seed germination and flowering timing. Vegetative, reproductive and seed dispersal phases are indicated by solid, dashed and dotted lines respectively. Variation in temperature during the seasons are indicated by a color scale, from red (warm) to blue (cool).

observed, which could be a way to increase chances of survival but could also result from the cohabitation of several genetically different subpopulations [11]. To regulate their life cycle, *A. thaliana* accessions developed different strategies. For example winter annuals require a long exposure to cold before to flower (vernalisation) so that in greenhouse winter annuals will be late flowering compared to summer annuals but flowering can be induced by leaving the plant at 4 °C for several weeks. On the other hand, summer annuals developed strategies to maintain seed dormancy during fall and some of them require long exposure to cold to germinate.

1.2 Genetic & epigenetic diversity

In this section, I give an overview of the large genetic and epigenetic diversity segregating within *A. thaliana* and show how the *A. thaliana* particular population biology and history as well as the environment influence this pattern of diversity (figure 5).

1.2.1 *A. thaliana* genetic variation

From the nineties to the early-2000s, genetic diversity in *A. thaliana* nuclear, chloroplastic and mitochondrial genomes was analysed using a variety of genetic markers, such as allozymes, RFLP, AFLP, CAPS, microsatellites or SNPs detected within specific genes (table 1). At this

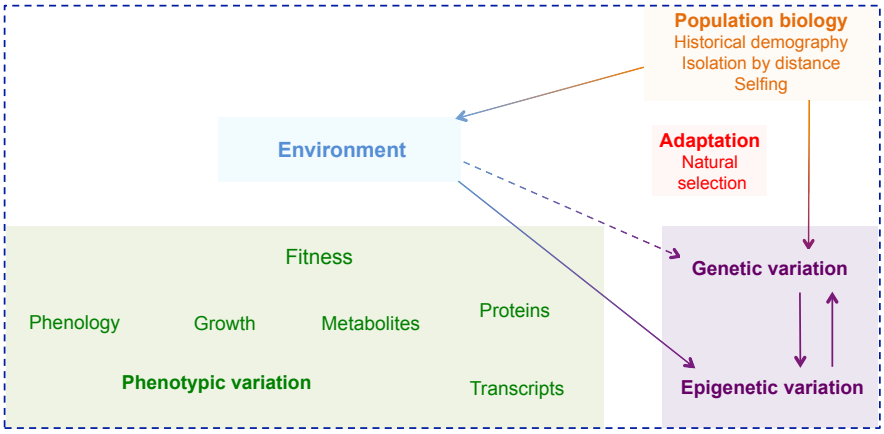


Fig. 5. ***A. thaliana* genetic and epigenetic diversity** is influenced by the demographic history of the species and by environmental parameters. Illustration of the topics explored in the subchapter 2 of chapter 1 of the introduction.

time, genotyping techniques and their cost were limiting the number of accessions and the number and position of genetic markers that could be genotyped. The considerable progress in genotyping and sequencing technologies that have been done during the last decade such as sequencing array technology and then sequencing technologies, allowed researchers to analyse more and more polymorphisms and more and more accessions so that now 1001 genomes of *A. thaliana* accessions are expected to be released in 2013, 450 being already available [12] (1001 Genome MPI, 1001 Genome SALK). Until now, nuclear genetic variation has been extensively studied in comparison to organelle genetic variation but the 1001 genomes project will help recovering organellar variants [13]. Table 1 lists some of the analyses that have been done regarding genetic diversity in *A. thaliana* and clearly illustrate the boom in the generation of genomic data that we have seen these last years.

1.2.1.1 Structure and demography

Overall, *A. thaliana* presents a high nuclear genetic diversity among and to a lesser extend within populations that can be surprising considering the selfing nature of the species. On the one hand, the low level of outcrossing that limits heterozygosity has been sufficient to create considerable local haplotypic diversity and genome wide LD has been essentially eroded by recombination on a very fine scale (5kb). On the other hand it has been low enough to generate quite structured populations at the worldwide and regional scales [4]. Cytoplasmic variability has been observed in *A. thaliana* but no relationship between the patterns of nuclear and cytoplasmic polymorphisms could be established probably due to the difference of nuclear and cytoplasmic transmssion over generation. Besides, the pattern of nuclear genetic diversity has been strongly influenced by the demographic history of the species and differs depending

Tab. 1. Analyses of genetic variation in *A. thaliana*.

Year	Reference	Type of markers	Marker location	M (a)	A (b)	P (c)	Geographic region
1989	Abott [20]	Allozyme	-	7	30	16	Great Britain
1994	Hanfsting [21]	CAPS	Nuclear	12	32	-	worldwide
		SNP	adh gene	12	37	-	worldwide
1996	Innan [22]	SNP	adh gene	75	17	-	worldwide
		Indels	adh gene	14	17	-	worldwide
1997	Kawabe [23]	SNP	chiA gene	120	17	-	worldwide
		Indels	chiA gene	24	17	-	worldwide
		Mononucleotide repeats	chiA gene	4	17	-	worldwide
1997	Innan [24]	Microsatellite	Nuclear	10	42	-	worldwide
1998	Purugganan [25]	SNP	Cal gene	91	17	-	worldwide
		Indels	Cal gene	13	17	-	worldwide
1998	Bergelson [26]	CAPS	Nuclear	3	115	11	worldwide
		CAPS	Mitochondrial	1	115	11	worldwide
1999	Miyashita [27]	AFLP	-	374 *	38	-	worldwide
2000	Sharbel [17]	AFLP	-	79 *	142	-	Eurasia
2000	Zwan [28]	Microsatellite	Nuclear	15	180	18	worldwide
2002	Barth [29]	CAPS	Nuclear	28	37	-	worldwide
		ISSR	-	41 *	37	-	worldwide
2003	Schmid [30]	SNP	Nuclear	8051	6-12	-	worldwide
		Indels	Nuclear	637	6-12	-	worldwide
2004	Jorgensen [31]	AFLP	-	131 *	95	53	North America
2005	McKhann [32]	SNP	10 nuclear gene fragments	197	95	-	worldwide
		SNP	4 nuclear gene fragment		265	-	worldwide
2005	Stenoien [33]	Microsatellite	Nuclear	25	104	10	Norway
2005	Nordborg [14]	SNP/ Indels	876 short nuclear fragments	17000	96	-	worldwide
2006	Ostrowski [34]	SNP	10 nuclear gene fragments		71	-	European
		Microsatellite	Nuclear	9	71	-	European
2006	Schmid [15]	SNP	Nuclear	115	351	-	worldwide
2007	Clark [35]	SNP *1 *2	Nuclear	648,570	20	-	worldwide
2007	He [36]	ISSR	-	165 *	560	19	China
		RAPD	-	162 *	560	19	China
2008	Beck [37]	SNP	Nuclear		475	167	worldwide
		SNP	Chlorosplastic		475	167	worldwide
		Microsatellite	Nuclear	8	475	167	worldwide
2008	Pico [38]	Microsatellite	Nuclear	16	175	7	Iberian Peninsula
		Microsatellite	Chlorosplastic	4	175	7	Iberian Peninsula
		SNP	Nuclear	109	175	7	Iberian Peninsula
2009	Montesinos [11]	SNP	Nuclear	76	188	10	Northern Spain
2010	Platt [4]	SNP	Nuclear	149	5707	-	worldwide
2010	Bomblies [5]	SNP	Nuclear	551	1005	77	Germany (Tübingun)
2010	Yin [39]	SNP/ Indels	11 chlorosplastic fragments	123	77	-	worldwide
2010	Moison [40]	SNP / Indels / Rearangements	Mitochondrial	15	95	-	worldwide
		SNP / Indels / Rearangements / Microsatellites	Chlorosplastic	68	95	-	worldwide
2011	Cao [19]	SNP *3	Nuclear	4,902,039	80	6 (2)	Eurasia
		Indels *3	Nuclear	810.467	80	6 (2)	Eurasia
2011	Gan [41]	SNP *3	Nuclear	3,070,000	18	-	worldwide
		Indels *3	Nuclear	1,200,000	18	-	worldwide
2012	Horton [18]	SNP *4	Nuclear	250,000	1307	-	worldwide
2013	Schmitz [42]	SNP *3	Nuclear	6,606,689	217	-	worldwide

(a) Number of polymorphic markers – (b) Total number of accessions – (c) Total number of 'populations'

* Total number of polymorphic bands. – *1 Non redundant SNP – *2 Array sequencing – *3 Illumina sequencing – *4 Array genotyping

on the region of the world considered (figure 6). In Europe and eastern Asia, the native range of the species, *A. thaliana* exhibits continuous isolation by distance at almost every geographical scale with low regional LD and low haplotype sharing [4, 14, 15, 16]. Based on spatial pattern of genetic variation, two regions, the Iberian Peninsula and the Caucasus region were likely the major refugias during the pleiostocene glaciation, 18,000 years ago, although other refugias could have existed [15, 16, 17]. After the glaciation, *A. thaliana* rapidly migrated from those refugias and admixed in Central and Eastern Europe. Populations from Northern Europe are quite differentiated compared to the population of Central Europe [18]. They are characterised by a higher haplotype sharing probably due to smaller population sizes. [14] The split between Northern and Central European populations probably took place quite early regarding re-colonisation of Europe by *A. thaliana*, before 7000 years ago and two different populations from the South and North East probably contributed to colonise Scandinavia [16]. Accessions from Northern Russia and Central Asia are more differentiated compared to the rest of Eurasian accessions. They are more uniform, show higher regional LD and present a higher number of deleterious mutations probably resulting from small population size [15, 19]. Overall this pattern of genetic diversity suggests that a small number of migrants recently and quickly colonised that region [15, 19]. Finally, *A. thaliana* has been introduced in North America in the course of human settlement over the last 300 years. Because a limited number of European haplotypes have probably been introduced repeatedly, North American populations are characterized by a reduced genetic diversity, genome wide LD and haplotype sharing [14]. Besides, several haplotypes are spread across the entire continent but at the continental scale, little or no population structure and isolation by distance has been observed. This pattern suggests that without established local population and so possibility of admixture, haplotypes can spread over great distances [4]. Finally, some analyses that focused on local scale genetic diversity suggest that there is considerable heterogeneity among sites separated from 50m to 50km, in size, genetic diversity and outcrossing rate [5]. Besides, some environments seem to favour genetic variation compared to others, such as coastal regions vs. mountains in the North of Spain and rural vs. urban habitats [5, 11]. Part of this variation can be attributed to variation in population size or pollinators populations that probably contribute to outcrossing.

1.2.1.2 The pattern of genetic variation

Beyond analysing population structure and demography, a huge interest of the sequencing of several accessions is the analysis of the pattern of genetic variation. In other words, how does genetic variation appear? How frequent are mutations? Where are modifications preferentially localised in the genome and what are there consequences on protein function?

First, the sequencing of mutation-accumulation lines (MA lines) derived from Col-0 for 30 generations allowed the estimation of spontaneous mutation rate to a value of 7×10^{-9}

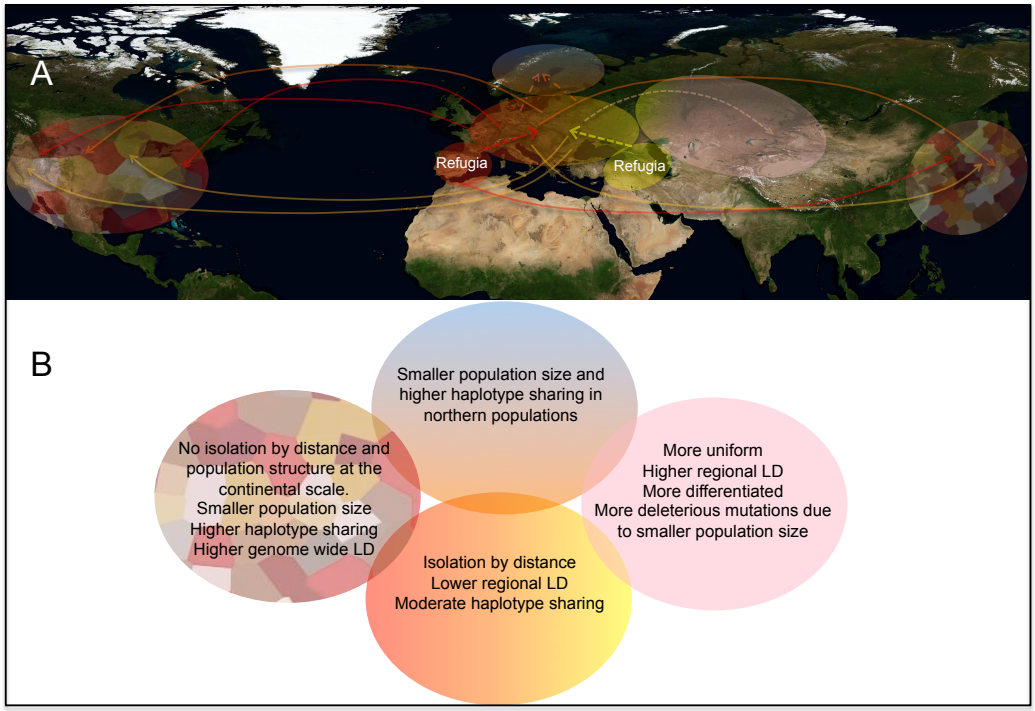


Fig. 6. ***A. thaliana* demographic history** (A) deduced from pattern of genetic diversity (B) observed within and among *A. thaliana* accessions. Arrows represent population migrations. Colors represent population genetic differentiations.

base substitutions per site per generation [43] (table 2). This estimation suggests that there is approximately one new mutation per seed produced during each round of selfing. As a consequence, it is not surprising that several mutations of varying phenotypic importance have been found in lab-propagated strains such as the mutation of the *ERECTA* gene in Ler seed stock. There is also approximately 0.6×10^{-9} and $0.3 \times 10{-}9$ chances of 1- to 3-bp deletions and insertions per site per generation respectively [43]. As expected, the mutation rate of dinucleotide microsatellite loci is much higher (between 2.3×10^{-3} and 4.96×10^{-5} mutation per allele per generation depending on the motif) [44]. The spectrum of base substitution mutation in MA lines is strongly biased. Indeed, over the 6 possible types of single nucleotide changes (figure 7), transitions are 2.4 times more frequent than transversion and G:C→A:T transitions account for approximately half of the observed mutations [43]. This phenomenon has been explained as the result of two main processes: the spontaneous deamination of methylated cytosine and the ultraviolet light-induced mutagenesis of dipyrimidine sites (figure 7, table 2). Because the MA lines have been done in Col-0 background and under greenhouse conditions, we can wonder if the spectrum of polymorphism observed in nature (where UV radiations are stronger) and under different genetic backgrounds is similar. In 80 accessions from Eurasia, the distribution of the different class also favours transitions over transversions. However among transitions, the bias of G:C→A:T was only observed for low frequency derived alleles but

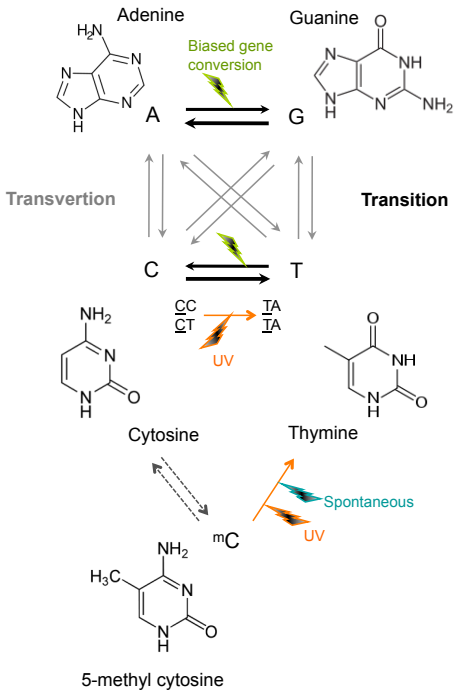


Fig. 7. Spectrum of base substitution mutations. In *A. thaliana*, transitions are favoured over transvertion mutations as represented by the size of the arrows. This spectrum partly results from environmental (orange) and non environmental (greens) factors.

not for high frequency alleles [19]. Thus this bias seems to be observed only for very recent mutations, which is also suggested by another study where they sequenced 13 genetically closely related accessions from North America (Becker, personal communication). This difference between old and recently acquired mutation could partially be explained by the effect of biased gene conversion [2]. Thus without considering demography and selection, environmental and non-environmental factors influence the pattern of polymorphism in *A. thaliana*. Overall, the genomes of two relatively distant accessions differ in average by 0.5%. at the species level and based on the Col-0 genome, there is more genetic variation in intergenic regions and transposable elements [41](table 2). However, a high proportion of mutations are also present in genes some of which lead to major changes of proteins structure such as modification of start or stop codons, premature stop codon, frameshifts and altered splice donor and acceptor sites [19, 35, 41, 43]. It is worth noting that, often, compensating changes have been identified thanks to the re-annotation of genomes and de novo-assemblies. For example, splicing variants can compensate splice site disruption and frameshifts can be restored by the combination of multiple indels [19, 41, 45]. As a consequence, genetic variation is not restricted to mutations but also to gene

2. Biased gene conversion is a neutral process associated with a recombination by which AT / GC heterozygote will produce a greater number of gamete carrying G or C than A or T through the GC-biased repair of A:C and G:T mismatch in heteroduplexed recombination intermediate) that favours G:C over A:T allele.

models. Regarding mutations affecting proteins, the distribution across gene families is not random. NB-LRR, F-box, RLK and RING families are predicted to encode highly variable proteins and are the most common ones represented in newly assembled fragments that did not align against the reference genome meaning that they are subjected to high copy number variation [19, 35, 41]. Finally, regarding the pattern of genetic variation in *A. thaliana*, there is a genome wide excess of rare polymorphisms [14, 15] that is likely due to the recent expansion of *A. thaliana* population but also to selective factors as there is a higher-than-expected proportion of non synonymous polymorphisms.

1.2.1.3 Large scale genetic variation

Most of the analyses presented so far regarding *A. thaliana* genetic diversity have focused on small scales variation i.e. substitution or insertions and deletions of relatively small size <100bp. Although small indels and deletions are much more frequent [45, 46], large scale insertions / deletions (>100bp) also segregate in population and significantly contribute to genetic diversity including genome size variation observed among accessions [47, 48]. The illumina sequencing of various accessions led to the identification of many reads that could be assembled *de novo* in contigs but could not be aligned with the Col-0 reference accessions (1-3Mb). Because some of them could be aligned to the genome of Ler accession [45] or to A. lyrata genome they were assigned as the ancestral state and so probably correspond to deletions in the Col-0 genome [19]. A large number of copy-number variants (CNVs) that likely corresponds to the duplication of a region but can also arise from more complex mechanisms, have also been detected, many of them being shared among accessions [19]. The distribution of large-scale insertions and deletions (including CNV) is not random as they are more frequent in pericentromeric regions than on chromosome arms, which could partially be explained by the mechanisms driving their appearance i.e. the activity of transposable elements (TE) and recombination-based mechanisms [49]. Interestingly, environmental constraints are known to influence the frequency of homologous recombination in somatic and meiotic cells [50] and several retrotransposons can be reactivated by biotic and abiotic stresses [51]. Genomic rearrangements could in fine be generated and lead to stress related phenotypic variation. For example, the progeny of tobacco plants challenged with a mozaic virus (TMV) present an increased frequency of rearrangements in the loci presenting homology with the LRR region of the gene of resistance to TMV (N-gene) [52], which could represent an attempt to make new TMV associated R genes. Another example is the exceptional accumulation of CNVs, overrepresented by TE and biotic responsive genes that can be observed in few generations under temperature and SA- mediated environmental stresses [53]. Finally, a strong variation in genome size attributed to variation in the number of 45s ribosomal repeat in *A. thaliana* Swedish accessions, was correlated with a North/South pattern of localisation of the individuals, which could suggest an adaptive role of

those copy number variations (M. Nordborg, personal commmunication). The exact molecular mechanisms at the origin of those observations are not clear but there are strong evidences that epigenetic modifications drive stress associated variations in homologous recombination frequency and transposons' reactivation [50, 51]. As a consequence, epigenetic variations could be an intermediate between the environment and stress induced genetic mutations.

1.2.2 A. thaliana epigenetic variation

Epigenetic variations refer to modifications of epigenetic marks such as post-translational histone modifications, histone variants, small or long non-coding RNAs and DNA methylation, that result in variation of chromatin condensation and accessibility and therefore affect transcription and genome stability. Those changes can be under environmental control and can be transmitted during mitosis and meiosis. Regarding natural epigenetic variation the most studied epigenetic mark is DNA methylation. In plants, cytosines in CG, CHG and CHH context (H being A, T or C) can be methylated (figure 7) [54, 55] and DNA methylation is predominantly found in centromers and pericentromeric regions that are enriched in transposons [56, 57] (table 2).

As for stress induced genetic mutation, an issue when studying epigenetic variation is the origin of epimutations and its complex relationship with genetic variation (figure 8). Regarding this problem, several classes of epivariants have been distinguished [58]. First, 'obligate' epimutations are completely dependent on a genetic variant whereas 'facilitated' epialleles are generated from a genetic mutation but can be maintained independently of this variant. Finally, 'pure' epialleles appear stochastically and completely independently of genetic sequence modifications (figure 8). In *A. thaliana*, most of the well-characterised natural epialleles belong to the two first groups. For instance, in Bor-4 accession, a deletion spanning the promoter region of *VIM1* gene that encodes a methylcytosine-binding protein is responsible for cytosine hypometylation of centromers. Because methylation is restored as soon as a functional *VIM1* allele is introduced by crossing in the Bor-4 background, methylation variation of centromeric repeats in Bor-4 is an 'obligate' epiallele [59] (figure 8). In the latter example, the genetic variation generating the epiallele was acting in trans. Some genetic mutations, such as the insertion of a transposon in the promoter or in the intron of a gene, can directly affect gene methylation in cis througt siRNA-based silencing mechanism [60, 61]. A different situation is illustrated by the complex genetic and epigenetic relationship between the *FOLT1* and *FOLT2* loci in *A. thaliana* [62]. Both loci contain a full-length functional copy of the *FOLT* gene which activity is important for proper folate transport and plant fertility. But *FOLT2* loci also contains multiple truncated copies that produce siRNAs able to target the functional *FOLT1* gene but does not reduce the expression of the functional *FOLT2* gene. In Col-0 accession, the $FOLT1[Col]$ gene is active and *FOLT2* locus is absent. In Sha accession, the two loci

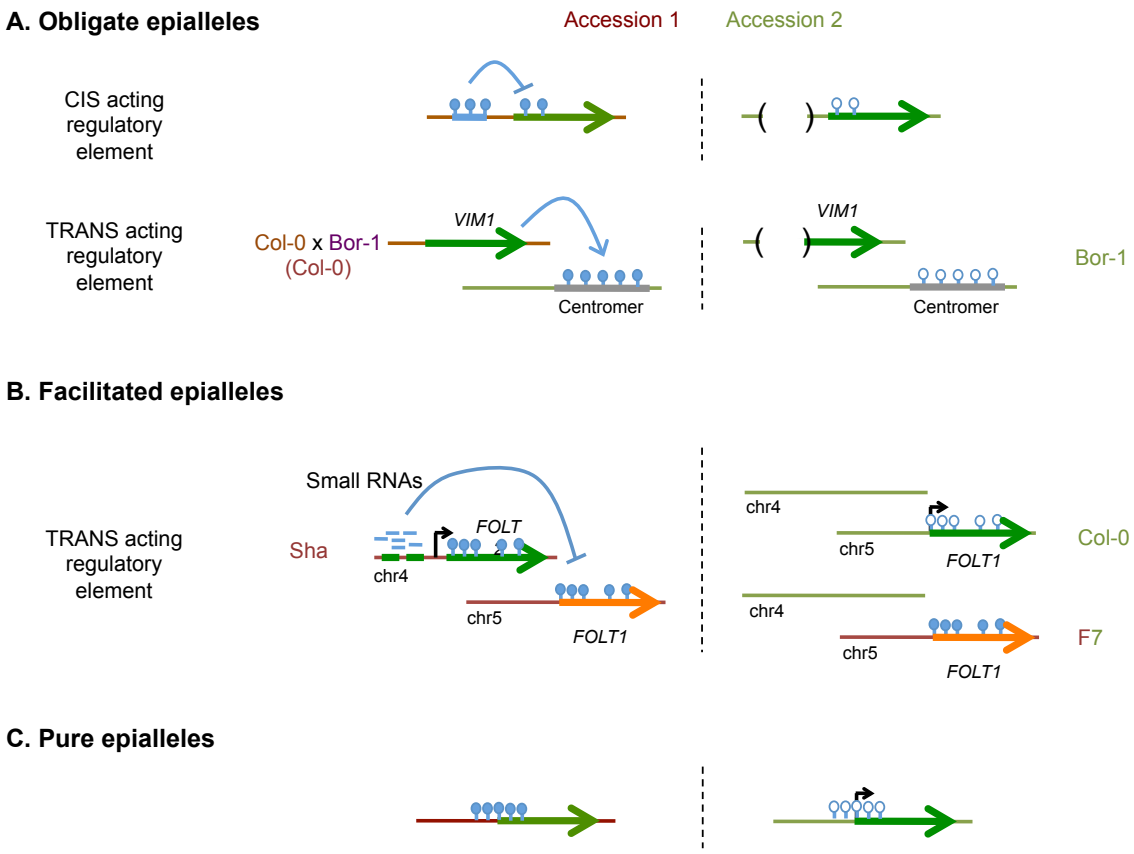


Fig. 8. The different classes of epialleles. 'Obligate' epimutations are completely dependent on a cis or trans genetic variants (A). If these genetic variations are removed, methylation disappeared. See the paper of Woo and colleagues 2007 for details about the *BOR1* example [59]. 'Facilitated' epialleles are generated from genetic mutations (B) but can be maintained independently of these variants. For example, the Sha allele of *FOLT1* remain methylated in a F7 lines carrying the Col allele at the locus *FOLT2* that does not produce the small RNAs that triggered methylation at the first place [62]. Finally, 'pure' epialleles appeared stochastically and completely independently of genetic sequence.

are present so that $FOLT1[Sha]$ gene is inactivated by the siRNA produced by $FOLT2^{[Sha]}$ locus but FOLT activity is maintained by the functional FOLT2 gene. Interestingly, F7 plants that have inherited the *FOLT2* region from Col-0 and the *FOLT1* locus from Sha lack FOLT activity because of the stably inherited DNA methylation at $FOLT1[Sha]$ gene, even though the inducing locus ($FOLT2[Sha]$) is not present anymore (figure 8). Thus, the methylated $FOLT1[Sha]$ is a 'facilitated' epiallele because it appeared from a genetic variant (likely an inverted repeat at $FOLT2[Sha]$ locus) but can be maintained independently from this variant [62]. Although a wide natural epigenetic variation is known to exist in *A. thaliana* [46, 63], the detection of natural 'pure' epialleles is particularly challenging because of the inherent genetic variation that is also segregating in accessions. Besides, without the complete information about the history of the genetic background accessions have passed through, it is impossible to know if an apparent 'pure' epiallele reached this state without any external genetic influence. The case of *FOLT* homologs illustrates particularly well this last point.

The MA lines (presented before p.26) are particularly prone to overcome this problem because their genetic background is much more homogeneous and their history known. The bisulfite sequencing of 3 and 10 MA lines that derived from a common ancestor for 3 and 31 generations respectively, revealed that although whole genome methylation patterns are largely stable and therefore heritable in *A. thaliana*, epimutation rate at single independent positions (DMP) is several orders of magnitude higher than genetic mutation rate (table 2). Contrarily, the epimutation rate of large contiguous methylated or unmethylated regions (>50bp, DMR) is relatively similar to the genetic mutation rate. Overall, 6% of methylated cytosines were detected as significantly differentially methylated in at least one of the ten 31st generation MA lines compared to the 3rd generation ones and this value is probably underestimated due to sequencing depth that leaves poorly covered regions and low statistical power to detect differentially methylated sites [64, 65].

As for genetic variation, the spectrum of epigenetic variation is far from random (table 2). First, although, most differentially methylated cytosine affect independent positions rather than large contiguous regions (>50bp), epimutations were preferentially observed in genes whereas methylation on or near transposable elements was much more stable [64, 65]. Interestingly, this pattern has also been observed among natural accessions [46, 63, 42]. The stability of methylation in transposons is not surprising as transposon's reactivation through demethylation can lead to transposition and so genetic mutations [66]. However, given that methylation variations in genes are only rarely associated to transcriptional changes [63, 64, 65], the origin and significance of this variation is less intuitive. One hypothesis is that epimutations in genes arise spontaneously and are then maintained imperfectly by MET1 (METHYLTRASNFERASE 1) as it has been suggested in mammals [67]. Then, certain positions are particularly prone to increase or decrease in methylation because the same changes in independent MA lines

Tab. 2. Comparing the general trends of genetic and epigenetic (DNA methylation) variants.

	Genetic (SNP)	**Epigenetic (DMP)**
Mutation rate	7×10^{-9}*1	4.5×10^{-4}*2
Preferential context	Intergenic regions and transposable elements	Gene bodies (CDS) & Intergenic regions
Spectrum	G:C→A:T and A:T→G:C more frequent	GC context more frequent (but see *3)
	Potentially all sites	A limited number of sites
Reversibility	Not reversible	Reversible
Selection	yes	?
Nature of variation	Qualitative and stable at the organism scale	Quantitative and variable at the organism scale

SNP: Single nucleotide polymorphisms – DMP: Differentially methylated position

*1 base substitutions per site per generation [43] – *2 methylation polymorphisms per CG site per generation [65] – * 3 The over-representation of GC among DMP could result from a higher statistical power in detecting a change at CG sites, which are on average much more highly methylated

were observed much more often than expected by chance [64] and a significant overlap exist between DMPs observed in MA lines and the ones observed in closely-related accessions from North America (personal communication, Becker). This observation combined to the fact that the epimutation rate estimated from the comparison of the 31st and 32nd generation MA lines was higher than the one estimated from the comparison of the 3rd and 31st generation MA lines suggests that a pool of methylated cytosines in the genome are particularly prone to epigenetic variation and meta-stably switched between low and high methylation states [64, 67]. This capacity of reversibility strongly differentiates epigenetic variation from genetic mutations that are typically considered unidirectional over evolutionary time (table 2). The stability and relatively low rate of genetic mutations over generations make them good markers to identify the phylogenetic relationships among distant individuals (i.e distant accessions). For that purpose, epigenetic variations are likely to be less efficient due to relatively high frequency and reversibility [42]. On the contrary, epigenetic variants might be much more efficient than genetic polymorphisms to analyse genealogies among closely related individuals (for example from a common location).

Whereas great variation of cytosine methylation has been observed among *A. thaliana* accessions, the other epigenetic marks such as post-translational histone modifications [68] and histone variants have been poorly surveyed. Besides, there are growing evidences that developmental and environmental stimuli including biotic and abiotic constraints can affect epigenetic marks. For instance, during vernalisation, the famous *FLC* locus is silenced via the progressive enrichment in tri methylation at the lysine 27 of histone 3 which in turn lead to the release of flowering induction programmes. Another example, observed in differentiated tissues after heat, freezing and UV-B stresses is the transient release of transcriptional gene silencing in several loci from centromeric and pericentromeric chromosomal regions via chromatin decondensation

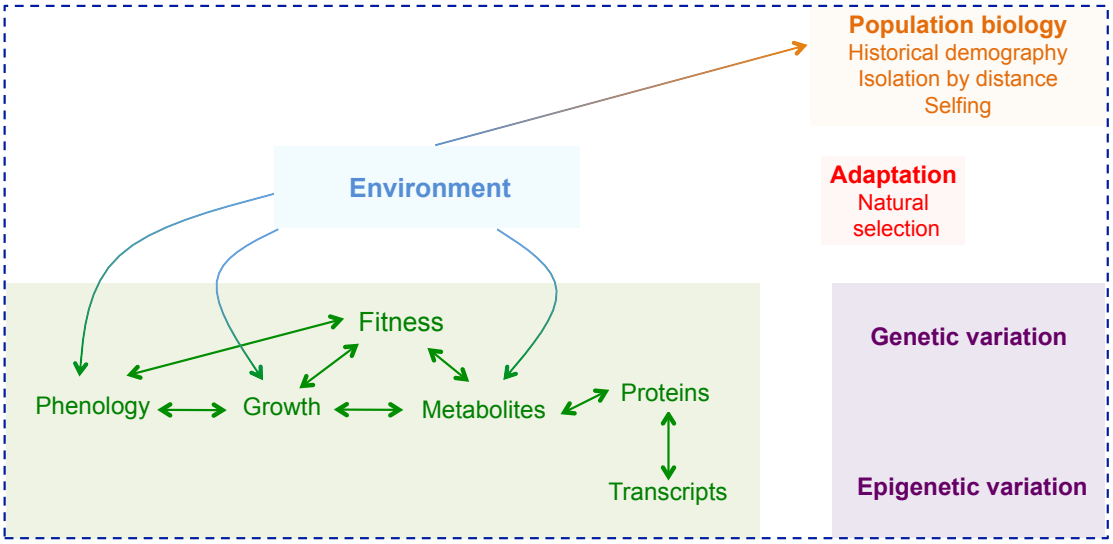


Fig. 9. *A. thaliana* phenotypic variation and environmental plasticity. Illustration of the topics explored in the subchapter 3 of chapter 1 of the introduction.

[69, 70]. In the two former examples, stress induced epigenetic variation is not transmitted to the progeny but transgenerational epigenetic changes have also been reported under various biotic and abiotic stress [71, 72]. Therefore it would be interesting to survey epigenetic natural variation in different cellular and environmental contexts and to link this variation to phenotypic variation. This last point is essential to understand if epimutations contribute to the adaptation of plants to their environment over generations.

1.3 Phenotypic diversity

In this section, I will focus on the description of tools available to analyse phenotypic variation in *A. thaliana*. Then I will show how phenotypic variation is integrated at different scales and how it is influenced by environmental conditions (figure 9).

1.3.1 Quantitative phenotypic variation

In theory, phenotypic traits can be separated in two different categories. First, qualitative traits display clear distinct phenotypic classes such as presence vs absence or white vs colourful. Most of the time, single genes explain the variation in those traits and the loci are relatively easy to map thanks to their clear mendelian segregation in the progeny of crosses (25:50:25) (figure 10). Quantitative traits are characterized by continuous phenotypic variation explained by several loci, named quantitative trait loci (QTL) that can contribute positively or negatively to the trait of interest and result in complex epistatic interactions (See 2.2.4). Besides, the environment often has a large effect on quantitative traits and the effect of loci may vary

depending on the environmental conditions. For example, flowering time is a quantitative trait governed by the action of several genes (*FLC*, *FRI*) whose action is strongly influenced by the temperature, day length and biotic or abiotic constraints the plant is facing [73]. In crosses, recombination reshuffles genes and the numerous resulting combinations of alleles create a wide phenotypic variation that cannot easily be associated to genotypic variation without a quantitative and statistical approach. Chapter 2 will focus on the different tools available to identify those loci responsible for quantitative phenotypic variation. It is worth noting that the distinction between quantitative and qualitative phenotypic variation is not so trivial and some traits could be classified as both. For example a protein can be phosphorylated or not (qualitative) but within a cell or a tissue the amount of phosphorylated protein can vary quantitatively. For most traits, it is possible to calculate the fraction of total phenotypic variance that can be attributed to genetic variance, the so called heritability. Heritability values mainly estimate the relative contribution of genetic variation to phenotypic variation compared to other non-genetic factors that are mainly environmental or stochastic. Phenotypic stochasticity can be caused by a lack of phenotyping accuracy, poor experimental designs or epigenetic variation (pure and maybe facilitated). As example, a meta-analysis of 107 phenotypes in *A. thaliana* natural accessions estimates the heritability of 74 traits as varying between 42 to 100%, the most heritable traits being the ones associated with flowering, development, ionomics and also defence-related phenotypes [74].

1.3.2 Phenotyping quantitative traits

Analysing quantitative traits through the phenotyping of hundreds to thousands of individuals requires precise high-throughput phenotyping techniques under tightly controlled conditions in order to reduce the phenotypic variance due to non-genetic factors (i.e. environmental variance, technical variance). Digital imaging strongly improved our ability to describe developmental parameters such as growth and shape because digital images can be analysed directly using softwares that automate the translation of digital data into mathematical values and their statistical analysis. As a result, different automated platforms and softwares have been developed for different traits. For example PHENOPSIS [75] and PHENOSCOPE [76] (figure 11) platforms that aim at analysing shoot growth under highly controlled conditions, GERMINATOR [77] for seed-germination phenotyping, LAMINA [78] and LeafPROCESSOR [79] to analyse leaf shape and Ez-Rhizo [80] or RootTrace [81] for root growth. Of course, high throughput phenotyping is not limited to developmental traits. Using Infrared thermography to measure leaf temperature and stomatal conductance or working with chlorophyll fluorescence (Growscreen Fluoro) [82] that captures the intrinsic photochemical efficiency of light harvesting in photosystem II, one can non-destructively and quantitatively assess the physiological status of various plants. Finally, large scales molecular detection techniques are also used to study

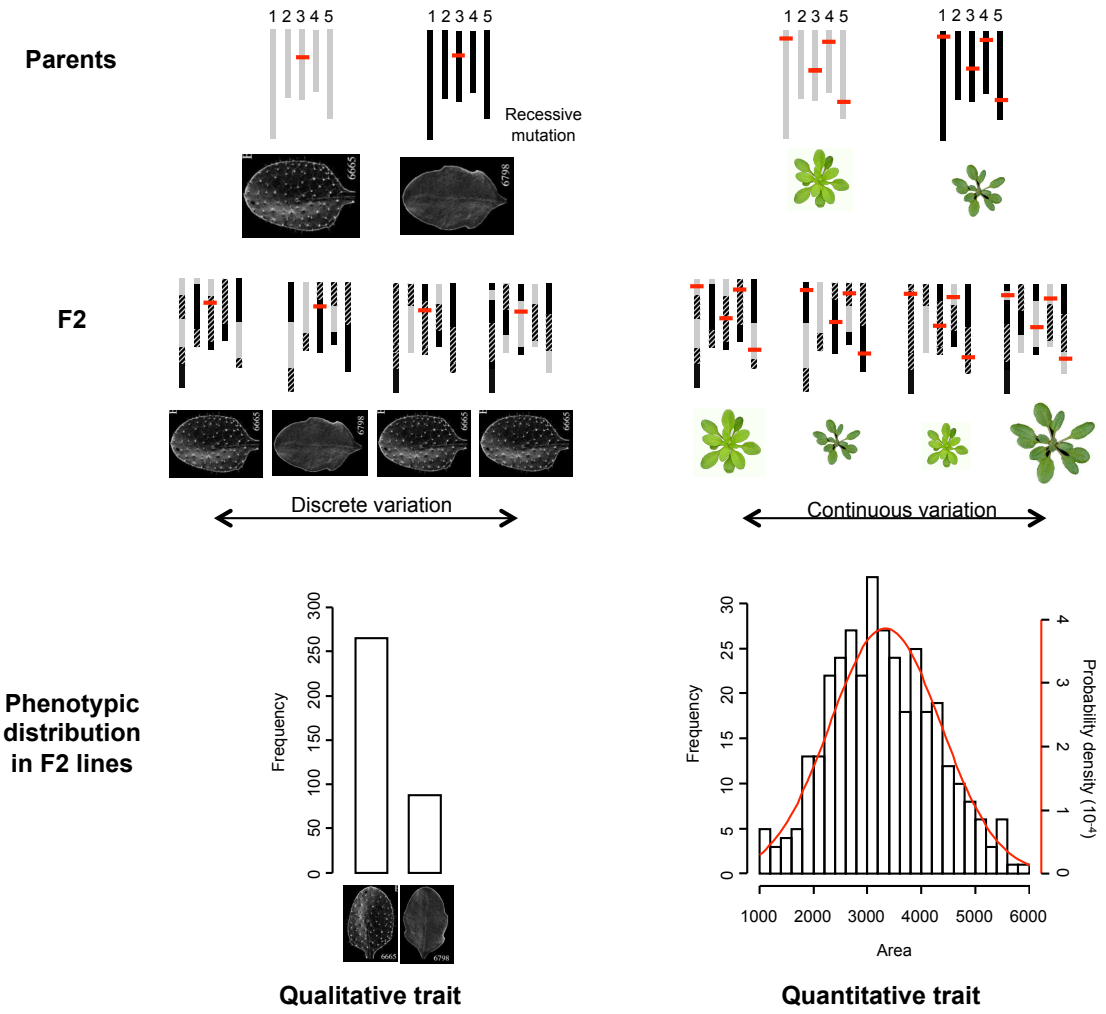


Fig. 10. Qualitative and quantitative phenotypic traits. The 'glabrous' phenotype (whether leaves are completely devoided of trichomes or not) is a qualitative phenotype whereas shoot growth is a quantitative trait. However, note that the number or density of trichomes can be quantitative phenotype in some populations. The position of the loci responsible for the phenotypic variation observed in the parents are indicated in red on the five chromosomes. The variants associated with those loci segregate in F2 population resulting in different patterns of phenotypic variation.

A. thaliana phenotypic variation. The transcriptome, metabolome, proteome and ionome of different *A. thaliana* accessions have been assessed using DNA microarrays or RNA-sequencing (transcriptome), gaz and liquid chromatography coupled with mass-spectrometry as well as nuclear magnetic resonance (metabolome, proteome, [83, 84], see [85] for detail description of those techniques) and inductively coupled plasma mass spectrometry (ionome, [86], see [87] for all the available techniques).

1.3.3 Integrated phenotypic variation and plasticity

The development of 'omics' techniques allowed the precise phenotyping of several accessions and mapping populations (see 2.1.1.1) at different levels of biological organisation and under various environments. As expected from the genetic variation observed among *A. thaliana* accessions, this species also exhibit extensive phenotypic variation. This phenotypic diversity is illustrated for rosette morphology on figure 11 but has been observed for many other phenotypic traits [88, 89] and at different scales i.e from gene expression to protein and metabolite accumulation, to growth and phenology.

Within all levels (transcriptome, proteome, metabolome, ionome, growth...) extensive heritable variation has been observed and significant correlations between various traits have been identified. These correlations revealed overall relationships between traits and so possible common genetic bases or functional relatedness. They also brought insights into the organisation of those biological pathways into networks. The relationships between the different levels of biological organisation have also been assessed using correlations and multiple regression analyses. For example, Meyer and colleagues showed that a specific combination of low-molecular-weight metabolites could be used to predict biomass [90] whereas in another study, starch, identified as the best integrator of plant metabolic status, was the best predictor of rosette fresh weight [91]. Finally, Prinzenberg et al. showed that ionomic traits might be the main drivers of several growth related traits [92]. Those apparently contradictory conclusions may be explained by differences in the populations under study, in the statistical methods used or in the environmental conditions. However, they may also highlight the complexity of growth-related phenotypic traits that integrate many different biological pathways. Similar levels of intricacy are expected for life history traits (flowering time and seed germination).

Plants are facing many changing and challenging environmental conditions during their life cycle to which they have to adapt using specific responses at the different levels of biological organisation. For example, transcriptomic profiles are largely affected during moderate drought stress [93, 94]. Strong constraints on metabolic profiles [95, 96], ion homeostasis [92, 97] and physiological parameters such as epidermal conductance and cell wall properties can be observed. At another scale, root development is favoured over shoot growth [98] (figure 11) and

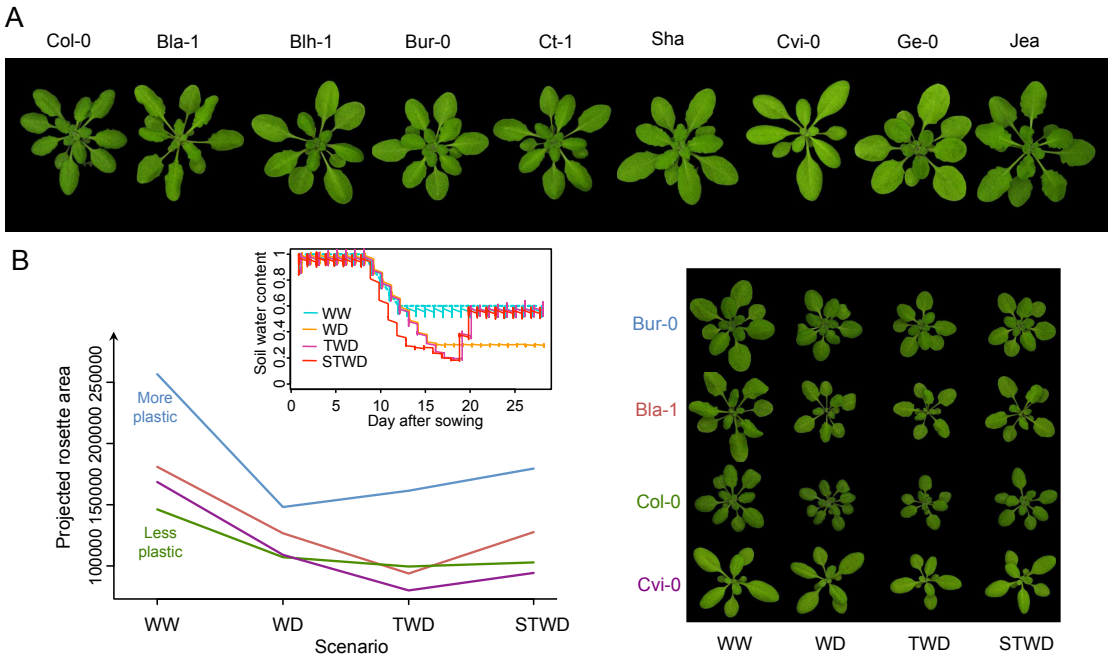


Fig. 11. *A. thaliana* phenotypic diversity illustrated on accessions for rosette related traits. A. Plants were grown for 28 days in short-day condition on the PHENOSCOPE in well watered conditions (WW). Note differences in leaf shape and flatness and petiole length. B. Projected rosette area reaction norms and pictures of 4 accessions grown on the PHENOSCOPE for 28 days in short-day condition under well watered (WW), water deficit (WD), transient water deficit (TWD) and strong transient water deficit (STWD) conditions. Note that Bur-0 and Cvi-0 show the higher and lower plasticity respectively in response to water deficit compared to well watered condition. This variation in plasticity (the so called GxE interaction) is observed on the graph by non parallel reaction norms. The evolution of soil water content associated with each environmental scenario are indicated. Results credit: Sébastien Tisné.

plant life cycle is often shortened in response to drought [99]. Plant phenotypic plasticity [3] does not only depend on the type of stress but also on the intensity of the stress. For example, a moderate drought stress does not result in the same phenotypic changes than an extreme one [98] (figure 11). It is important to note that all phenotypic changes that arise in response to environmental perturbations are not active responses resulting from a specific signal perception-transduction system. Some of them might be the results of passive physiological or developmental constraints or arise from genetic correlations with other traits [100]. Among Arabidopsis accessions, wide variation in phenotypic plasticity has been observed [89, 101] (figure 11) by testing genotype x environment interaction using a standard analysis of variance [4]. Besides, correlation analyses have been performed in order to identify coregulated genes, proteins, metabolites, ions and growth parameters in response to environmental constraints. Overall, correlation analyses can reveal causal relationships that have then to be tested experimentally but don't give any information about the underlying genetic bases if they exist. Indeed, correlation may reflect developmental or spatial control of two independent variables or may even be spurious [102]. Finally, using several methods of investigation such as global principal component analysis, one can define groups of individuals according to their responses to a specific constraint. Those studies allow the identification of extreme accessions that could be the most suitable to set up segregating populations and investigate the genetic causes of response variation.

A better understanding of phenotypic integration [5] in response to environmental consraints is important to predict the evolution of a population and how new genetic variants will be integrated at the phenotypic level under various environments.

3. At the individual level, phenotypic plasticity is the ability of a genotype to express different phenotypes across environments throughout its ontology. At the population level, phenotypic plasticity refers to the variation of traits' population mean across environmental conditions.

4. Analysis of variance typically partitioned the variance of a phenotypic trait as follow: Vp = Vg + Ve + Vexg + Verror where Vp corresponds to the total phenotypic variance, Vg, Ve and Vgxe to the proportions of phenotypic variance explained by genetic, environmental and plasticity conponents respectively and V error to the unexplained genetic variance

5. Phenotypic integration refers to the biological relationships among multiple subtraits of a complex phenotype and their relationships with other levels of biological organisation in a given organism – See [103]

2. FROM PHENOTYPIC VARIATION TO MOLECULAR VARIANTS AND VICE VERSA

In the last chapter, I described the extent of genetic, epigenetic and phenotypic diversity observed in *A. thaliana*. Part of this phenotypic variation can be explained by genetic and epigenetic modifications but not all mutations are responsible for phenotypic diversity. A long standing question in quantitative genetics has been to identify among all the variants observed between accessions the ones responsible for phenotypic variation (figure 12). In this chapter, I will describe the methods available to identify and validate the genes (QTG) and polymorphisms (QTP) underlying QTLs. Besides, through several examples, I will present what quantitative genetics teach us about the complexity of the genetic architecture of natural phenotypic variation. Finally I will bring some insights of the importance to combine all the information obtained from genomics approaches for various traits and under different environments into more global networks in order to better understand the consequences of genetic and epigenetic variation on the plant phenotype as a whole.

2.1 Mapping and cloning the genes responsible for phenotypic variation

Overall, the mapping and cloning of quantitative trait loci is based on the association of the phenotypic and genotypic variation observed within a group of individuals usually from the same species that can be more or less related to each other. The goals of such associations are to detect QTLs (and, potentially interaction among QTLs), to locate them on chromosomes with the best possible precision (the best being to identify the causal polymorphisms) and to estimate their effects. If these associations are performed on different environments the environmental specificity of QTLs can also be assessed. The power to detect significant associations depends on the genetic architecture of the trait of interest, i.e. the number of loci responsible for trait variation, the relation between the different loci and their effect, as well as trait heritability and the mapping difficulties associated to trait genetic architecture with be discussed in the second section. Besides, because genetic variants are reshuffled within a population thanks to recombination events during meiosis, the type of population (experimental cross design and accessions) and the number of individuals analysed can strongly influence the power of QTL mapping. Finally, the number, type and distribution of genetic markers that allow the

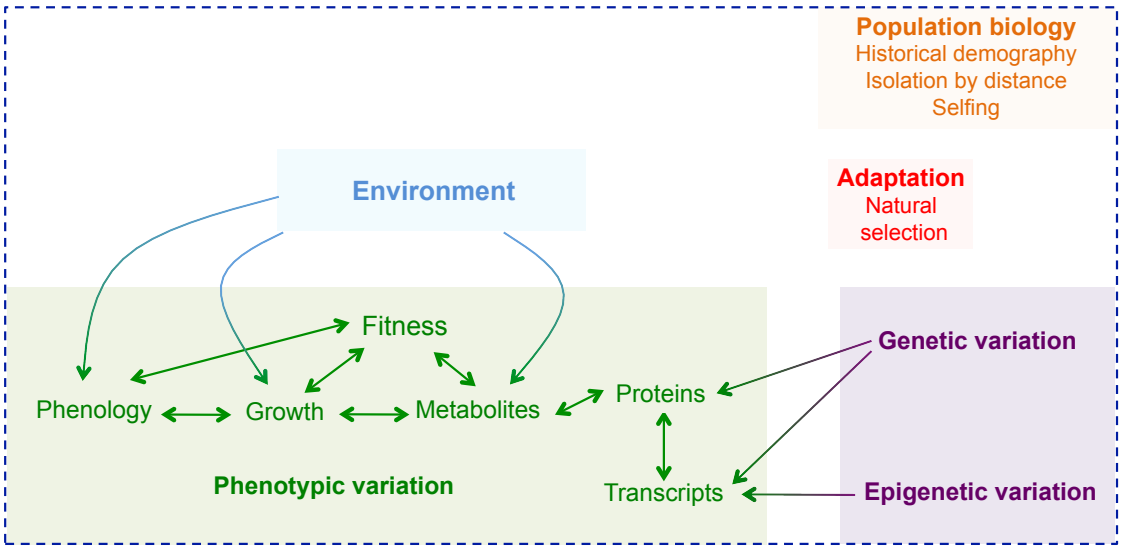


Fig. 12. The genetic and epigenetic causes of natural variation. Illustration of the topics aborded in the chapter 2 of the introduction.

identification of recombination events and the definition of QTL intervals will also affect the output of the mapping analyses. Different strategies can be used to identify QTLs and the choice of the strategy to conduct depends on the trait of interest, the genetic resource available and the biological question addressed. Indeed some researchers are mapping QTLs in order to identify new genes involved in a specific pathway whereas others will be more interested in the evolutionary and ecological importance of the underlying QTLs, and others will be essentially looking for markers to use in selection. In the first case, the cloning of at least one gene would be satisfying whereas in the second and third cases the number and effects of the QTLs in the whole population are the key results. In this section, I will present those strategies and start to highlight their strengths and drawbacks.

2.1.1 Linkage mapping in segregating population

2.1.1.1 Mapping populations

To map QTL, plant geneticists traditionally perform linkage mapping analyses in experimental segregating populations issued from the crossing of a limited number of accessions (figure 13), a method I will call 'linkage mapping' in the rest of the manuscript. Parental accessions may or not vary for the trait/genotype of interest but the F2 progenies or later generations have to show significant phenotypic variation to be able to associate it with segregating genetic markers. The different types of segregating populations available in *A. thaliana* are presented in figure 13. F2 populations have the advantage to be generated in few months but because many heterozygous regions are segregating in those populations, many lines need to

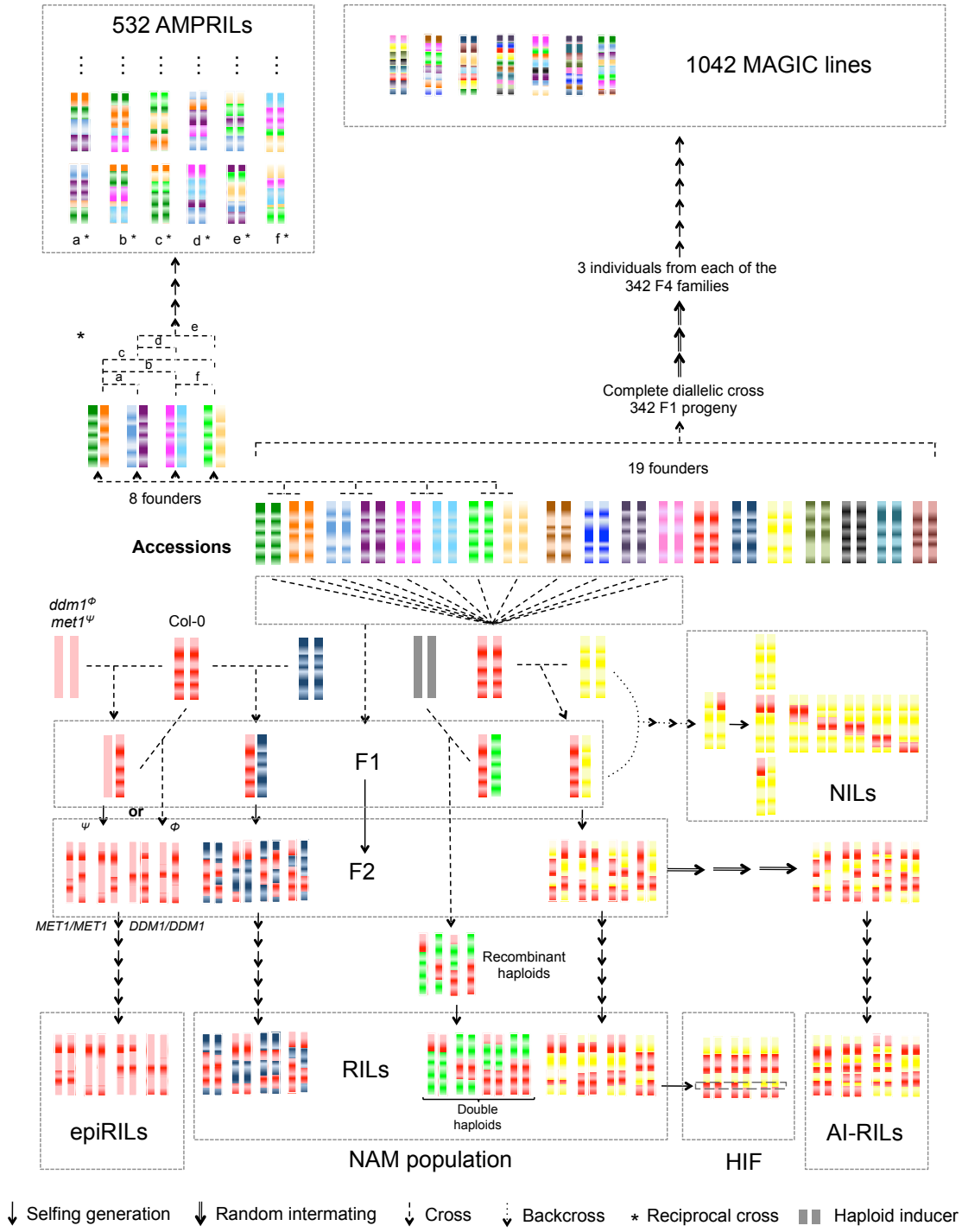


Fig. 13. Segregating populations for mapping quantitative traits in *A. thaliana*. The different colors represent genetic variation among accessions. Only one chromosome pair is represented per individuals, the darker regions representing methylated alleles and the lighter ones demethylated alleles. EpiRILs are a specific form of induced variation. RILs: Recombinant inbred lines, epiRILs: epigenetic recombinant inbred lines, AI-RILs: Advanced intercross-recombinant inbred lines, HIF: heterogeneous inbred family, NILs: Near-isogenic lines, AMPRILs: Arabidopsis multiparent RILs, MAGIC lines: Multiparent advanced generation intercross lines. This scheme has been simplified to allow the representation of all mapping populations so that: AMPRILs and MAGIC lines have actually only two founders in common.

be phenotyped to obtain a good power for QTL detection. Besides individuals have to be genotyped at each experiment and can't be phenotyped repeatedly. As a consequence, *A. thaliana* community has invested considerable efforts to produce recombinant inbred lines (RILs) populations (figure 13). Those populations are generated from F2 lines by repeated selfing and consist in series of nearly homozygous individuals that each represents a unique mosaic of the parental genomes [104, 105, 106, 107]. As a consequence of homozygosity, each RIL need to be genotyped only once but can be phenotyped repeatedly in the same or different environmental conditions and for different traits, which allows a precise and complementary phenotyping of each individual line. Over 60 bi-parental RILs populations are available in stock centers (1) and this number will probably increase now that their long generation time (7-8 generations) has been circumvented by the creation of doubled haploid lines (figure 13) [108, 109]. A drawback of using two-way RILs populations for QTL mapping is the poor mapping resolution (usually from hundreds of kb to several Mb) that is partially due to the limited number of recombination events observed. This number can be further increased in advanced intercross RILs (AI-RILs; figure 13) in which the F2 lines undergo several generations of intermating before fixation through inbreeding [110]. However a direct comparison of both RILs and AI-RILs power to detect QTLs has not been directly performed. In F2, RILs and AI-RILs populations, only two parental accessions are used, which strongly limits the genetic diversity segregating in these materials compared to the whole *A. thaliana* genetic diversity. To overcome this issue, two populations were generated from multiple parental individuals. In the Arabidopsis multiparent RILs (AMPRILs) the genomes of 8 parental accessions are segregating in 6 (12 if we consider the reciprocal crosses) independent RILs sets, each of them being the mosaic of 4 founders (figure 13) [111]. In the multiparent advanced generation intercross lines (MAGIC) the genomes of 19 parental accessions are segregating, one line being on average the mosaic of 9.97 distinct founders (figure 13) [112]. Because they capture more of the genetic and phenotypic diversity segregating in *A. thaliana*, AMPRILs and MAGIC lines are expected to allow higher power for QTL detection. However, some of the QTLs identified in diallelic crosses, could not be found with these more complex populations. An explanation for that is likely to be the complexity of the genetic interactions between the many different alleles at different QTLs. In the MAGIC lines, this complexity also makes difficult the detection of epistatic interactions between different loci. Besides, if a QTL allele occurred in just a single founder, its effect is diluted in the whole population resulting in a lack of power to recognize the effect of this specific allele. Regarding the precision of QTLs mapping, with a fixed population size, the population that accumulates more recombinations (i.e. MAGIC followed by AMPRILS, RILs and finally F2) generally results in a better resolution. For typical experiment size, RILs define QTL intervals of several Mb whereas MAGIC lines increase resolution to hundreds of kb.

Finally, more and more researchers start looking for the epigenetic variants responsible for

A. thaliana phenotypic variation. The mapping population described above segregate for both genetic and epigenetic variation making them not specifically suited for this goal, however any epiallele that remains perfectly linked to a genetic allele will be taken into account and its potential effects can be mapped in the segregating population. Another type of genetic material uses induced epigenetic variation instead of natural variation: two of these epigenetic recombinant inbred lines (EpiRILs) populations (figure 13) with limited genetic but high epigenetic variability have been developed [113, 114]. Theses populations are derived from an F1 issued from the cross between the Col-0 background and a mutant in the same background carrying a defective allele at either *MET1* or *DDM1* (*DECREASE IN DNA METHYLATION 1*). MET1 is a methyltransferase essential for the maintenance of CG methylation in repeated sequences and gene bodies whereas the chromatin remodeler DDM1 is involved in the maintenance of cytosine methylation in all contexts (CG, CHG, CHH) but mainly in repeated sequences [113, 114]. Both met1 and ddm1 mutants show a strong reduction of overall DNA methylation that can be relatively stably transmitted over generation even if the WT copy of the gene has been reintroduced. This allowed the fixation through backcross of different methylated or demethylated regions along chromosomes that have been reshuffled by recombination in F2 lines selected for *MET1/MET1* and *DDM1/DDM1* alleles (figure 13). Thus each resulting epiRIL is a mosaic composed of large chromosomal fragments inherited from the demethylated mutants or WT ancestors. Interestingly, demethylation of a significant proportion of loci was reverted likely as the result of de novo methylation rddm (RNA-directed DNA methylation) pathway whereas some WT methylated loci get demethylated after met1 cross. Besides, some of the demethylated transposons inherited from the met1 or ddm1 mutants where able to transpose during epiRILs formation resulting in several genetic variants in the final epiRIL population. Finally, heritable phenotypic variation, similar to what can be observed in natural *A. thaliana* accessions, have been observed for several traits within the epiRILs [113, 114, 115, 116] suggesting that heritable epigenetic variation could contribute to phenotypic differences in the wild.

2.1.1.2 Statistical methods

There are two main statistical problems associated with QTL mapping in experimental segregating populations. First, because the genetic information available in each inbred line is limited to specific marker loci along chromosomes, the genotype at intervening positions (including QTLs) has to be inferred from marker genotype data using a recombination model (missing data issues). Then identifying the best model, i.e how QTLs combine together and with other covariates to produce a given phenotype remains particularly challenging (model selection problem). The first and simplest method used for QTL detection (referred to as marker regression) consists in comparing the phenotypic distribution of individuals grouped according to their genotype at a given marker using a t-statistic or an ANOVA test. Because of its

simplicity, this method does not require genetic map information nor any specific software and can easily incorporate covariates. However it poorly estimates QTL effects and locations and have a low detection power when markers are widely spread [117]. A significant improvement over this method is interval mapping (IM) that takes into account missing genotype data at positions between pairs of consecutive markers [118]. Basically, this model uses the information from two adjacent markers to infer the likelihood of a QTL at any position between them [117]. More than dealing with missing genotype data, IM better estimates QTL effects and provides a LOD [6] curve that indicates evidence for QTL location [119]. Both marker regression and IM deal badly with closely linked QTLs. Including markers close to major-effect QTLs as covariates in composite interval mapping analyses (CIM) improves the ability to detect further modest effect QTLs by reducing residual variance [120, 121, 122]. Finally, because the effect of some QTLs depends on the genotype at other loci (what is called epistasis, cf below), some QTLs are not detected using one-dimensional genome scan analyses (marker regression, IM and CIM). Using two-dimensional genome scans allows the detection of these espistatic interactions and also increases the detection of linked loci and modest effect QTLs [123]. Another method consists in modelling multiple QTLs and their interactions simultaneously in a single analysis. Because all the possible models i.e. the number and position of QTLs and their interactions can't be analysed, the main problem of multiple QTL analyses is the search of the model that is best supported by the data. Model search can be done step by step either by adding markers one at a time in a simple model (forward selection), by removing them one by one from a large model (backward selection) or by alternating between both addition and deletion (stepwise selection). At each step, the new model is fitted to the data and compared to the previous one. The various methods developed for single scans (markers regression and IM) have been extended to the case of multiple QTLs (multiple regression [124] – Multiple Interval Mapping (MIM) [125]) and are used to fit models to the data. Overall, multiple QTL models probably fit better the reality of the complex genetic architecture underlying quantitative traits however the number of QTL probably remains underestimated mainly due to linkage and sample size [126].

2.1.1.3 Conclusion on the use of linkage mapping

Overall, in *A. thaliana*, linkage mapping allowed the identification of hundreds if not thousand of loci along chromosomes contributing to the variation of many different phenotypic traits [88, 89]. The requirement for only a few genetic markers to perform a complete genome scan has long been an important advantage of linkage mapping when genotyping was really expensive

6. The LOD score is the log10 likelihood ratio comparing the hypothesis that there is a QTL to the one that there is no QTL segregating at position z in the experimental population. Large LOD scores indicate higher evidence for the presence of a QTL.

(figure 14). Now the main advantage of linkage mapping is probably the power to detect rare variants, which increases the possibility of finding new alleles affecting genes involved in phenotypic variation but not identified in mutant screens. Besides, linkage mapping is performed in populations with known pedigrees and so is not subject to the effect of population structure ((figure 14), see details in GWA section 2.1.3). One of the main drawbacks of linkage mapping is the resolution of QTLs confidence intervals that often contains hundreds or thousands of genes when using standard RILs [127] (figure 14). As a consequence finding the genes and polymorphism responsible for phenotypic variation requires extra steps usually referred to as map-based cloning strategies that can be long and laborious.

2.1.2 Cloning strategies

Once a QTL interval has been identified, the segregation of a QTL must be individually confirmed in Heterogeneous Inbred Families (HIFs) or Near Isogenic Lines (NILs). HIFs are generated thanks to the residual heterozygosity in F6 or F7 RILs. The fixation of the heterozygous region observed in a given RIL allows the creation of a family composed of three different genotypes sharing a homozygous (but heterogeneous) genetic background except at the locus of interest that is either heterozygous or fixed for the parental alleles (figure 13) [128, 129]. NILs are generated directly from F2 by repeated backcrosses and contain a relatively small homozygous introgressed fragment from one parent in the isogenic background of the other parent [130, 131] (figure 13). The size of the introgressions depends on the number of backcross generations performed to obtain the NILs population [127]. First, the use of HIFs and NILs allows the conversion of QTLs into single mendelian factors minimizing the effect of residual genetic variation. Besides, the possibility to phenotype larger uniform progenies can help reduce environmental variation and give a better estimate of QTL effects. Finally, the phenotyping of heterozygous individuals also gives information about QTL dominance that was not obtained when studying RILs. It is worth noting that for a given QTL interval, several unique HIFs can be generated. Because of their heterogeneous genetic background, the HIFs may reveal different QTL effects and all of them may not segregate for the QTL, particularly when several QTLs with additive and epistatic effects segregate in the parental RILs populations. In NILs, the homogeneous genetic background limits combinations, but QTL effects and segregation can still vary depending on the parental genome. However different QTLs can be easily placed in a common genetic background using NILs (by crossing), which is not the case using HIFs. Finally, NILs population which introgressions cover the whole genome can also be used to map QTLs and generally allow the identification of smaller effect QTLs but with lower resolution than RIL populations [130, 132, 133]. Once a QTL has been confirmed in one or several HIFs or NILs, these lines are a good starting material for QTL fine-mapping. Briefly, the segregation of the QTL is tested in the descendants of the HIF or NIL that have

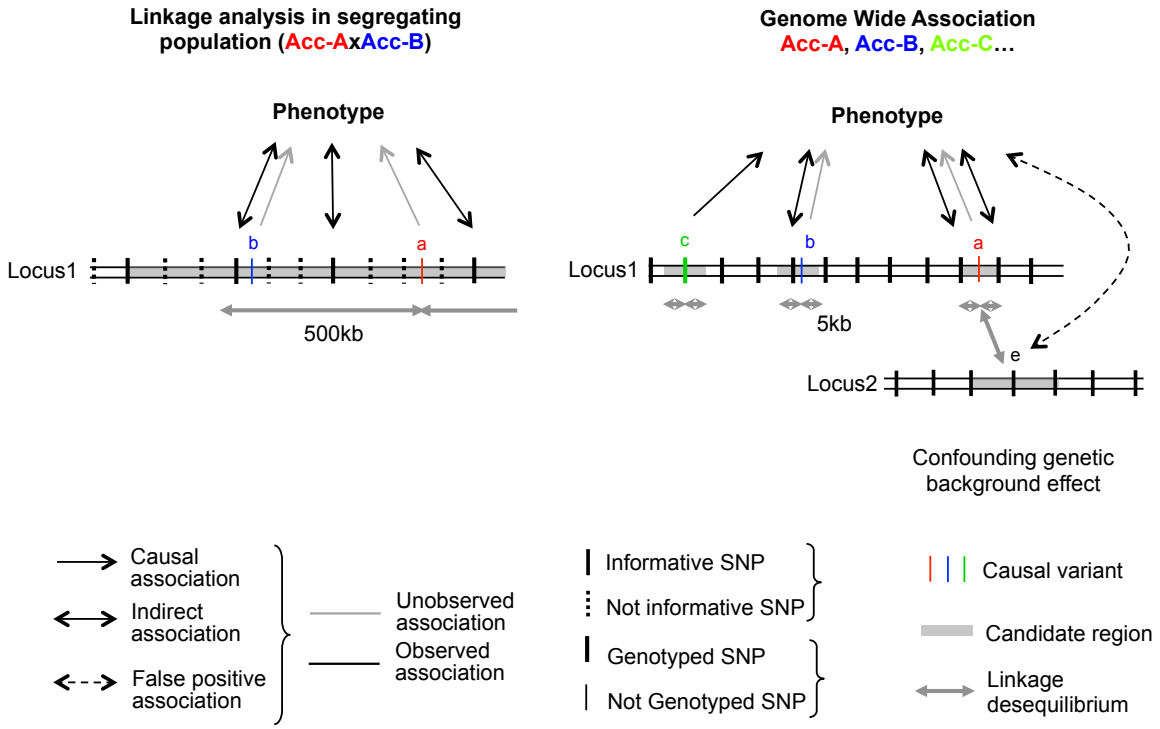


Fig. 14. Representation of Linkage and Genome Wide Association mapping advantages and drawbacks regarding the genetic parameters of mapping populations and accessions sets. Linkage disequilibrium is smaller in accessions sets compared to mapping populations due to a higher accumulation of recombinations over hundreds of generations. As a consequence, more SNPs are necessary to map recombination events in accession sets compared to mapping population. With the same genotyping data, many polymorphisms would be uninformative in a linkage mapping analysis. Beside, the power to detect associations and the maping resolution are higher in a GWA than in a linkage analysis (the QTL a and b can't be separated in linkage analysis). On the other hand, linkage analysis are less sensitive to confounding genetic background effect; e is a false positive association identified by GWA due to the genetic background of accessions ('spurious' linkage desequilibrium) . Because more alleles are segregating in accessions set, more association can be detected (c). Using 1001 genomes data could allow GWA to directly identify causal polymorphisms (c).

recombined within the heterozygous region in order to refine the interval supporting the QTL. As the interval is refined, more and more progenies need to be genotyped to identify informative recombinants and new markers are developed to delimit recombination breakpoints within the region of interest. The efficiency of map-based cloning to reduce the candidate interval to few genes will depend essentially on the frequency of recombination within the candidate interval as well as the genetic architecture of the trait within the QTL interval. Indeed, the phenotypic variation observed within an HIF can result from the additive contribution of several linked loci of small effect that could not be detected individually, so that the QTL can be lost while it is refined in recombinants if underlying loci's individual effects are too thin.

Another possibility to quickly identify the gene of interest consists in looking for mutations predicted to change genes/proteins properties within QTL intervals. The 1001 genomes sequences (see 1.2.1) will undoubtedly help for this purpose and the 1001 proteome website allows the direct visualisation of all the amino acid changes observed in the accessions sequenced by the MPI and JGI [134] (1001 proteomes). This method can work if the causal mutations are really obvious (stop codons or frame changes in genes previously identified as contributing to the phenotypic trait under study [135]) and/or when few genetic variants segregate within the candidate interval. However if the candidate QTL interval contains hundreds of genes and the responsible SNP is a non-synonymous mutation or lies in a cis-regulating sequence segregating with many others then it is almost impossible to guess the causal variant. More insights can be obtained from all the genomic data available in *A. thaliana* such as gene annotations (including predicted functions and gene ontology), protein interactions, gene expression data (in various tissues, under different environments and in different accessions (gbrowse of 19 MAGIC parental lines [41], Salk gbrowse [42]), genes coexpression... [136, 137, 138] . The association of those genomic data for all the genes within a candidate interval into networks can lead to the identification of interesting candidate genes [139].

2.1.3 Genome-Wide Association mapping

As opposed to 'linkage mapping' which is based on the use of experimental populations, genome-wide association (GWA) mapping exploits accessions' shared ancestry to directly identify the variants responsible for phenotypic variation. Because of the higher number of generations separating individual accessions compared to individual RILs or F2s, much more recombinations accumulated in natural populations. In *A. thaliana*, linkage disequilibrium (LD) [7], decays rapidly within 10kb so the optimal number of SNPs necessary to indirectly take into ac-

7. Two alleles are in linkage disequilibrium when they are not segregating independently in a population. Patterns of segregation are strongly dependent on the recombination rate than can vary along the genome. Overall, linkage disequilibrium concerns genetically close alleles (< 10kb in *A. thaliana*) or alleles that are located in regions with a low recombination rate (pericentromeric region). Nevertheless allelic incompatibilities can also result in long-range linkage disequilibrium.

count (through linkage) any polymorphism in the genome has been estimated between 140,000 and 240,000 [140] (figure 14). As a consequence, a 250k chip has been designed based on the result of the array sequencing of 19 accessions and used to genotype 1,307 worldwide accessions [18, 140]. The 1001 Genome project will probably improve GWA analyses by including new accessions and new variants in GWA projects and by increasing the chance to detect directly–or via mutations' effect predictions–the responsible variants [88].

Several studies showed that GWA mapping could identify many true genotype-phenotype associations in *A. thaliana*. Among others, a meta analysis performed on 76 to 194 accessions for 107 phenotypic traits related to flowering time, defense, ionomics and development, identified several associations previously mapped or cloned by traditional linkage mapping and cloning [74, 141, 142] such as *MOT1* [74, 143, 144], *HKT1* [74, 145, 146] and *ACD6* [74, 147] as well as new regions containing no a priori candidate genes. One of the great advantages of GWA over QTL mapping is the resolution power that can often be reduced to few kb i.e. a small number of genes (figure 14). Nevertheless some genomic regions show clusters of significant SNPs that are likely due to extended LD [74, 148] and can make the identification of candidate genes difficult. Also, clear spurious associations several tens of kb away from the real causal polymorphism also start to be described in the community. Concerning GWA power, another advantage that was not expected from human GWA analyses is the relatively low number of accessions that need to be included in the analysis at least to start to get significant associations. Atwell and colleagues showed that 96-192 accessions could be sufficient but this number actually depends on the architecture of the trait of interest and probably reflects the occurrence of common segregating alleles in the whole population [74, 127, 148].

Among the drawbacks of genome-wide association mapping, the one that has received the most attention is the confounding effect arising from populations' genetic background. Population structure [8] which generally describes remote common ancestry of large groups of individuals, can cause LD between the causal variants of a given trait and unlinked loci throughout the genome leading to spurious genotype-phenotype association (figure 14). This occurrence of false positives is particularly important when phenotypic variation overlaps with the pattern of population structure and/or environmental clines (because *A. thaliana* population structure is partially associated with geographic distances (at least in Eurasia, [14])). Finally, correlations between genetic variants can also be observed in 'unrelated individuals' although to a lesser extent [149]. Several methods have been used to take population relatedness into account in genotype-phenotype association studies. For example, methods using the cluster membership (Q) obtained with STRUCTURE software [150], or the coordinates of the first axes of principal-component analysis, as covariates in the association model perform well when popu-

8. Population structure is observed when a population can be divided in different subpopulations based on shared differences in genome-wide allele frequencies. Population structure often results from variation in the relatedness between individuals and can indicate relatively low migration rates between subpopulations.

lation structure is simple. An additional improvement, based on Fisher's observation in 1918 that the more alleles individuals share, the more similar they will be, models phenotype with a linear mixed models that accounts for the phenotypic variation that is linked with accessions pairwise relatedness [151]. Overall, these methods reduced the number of false positives but in the same time reduced sensitivity [74]. The apparition of false negatives when correcting population relatedness is especially likely when the environmental conditions that constraints phenotypic variation and genetic variants overlap with population structure. Besides, these models can artificially increase the association score of rare alleles so that they often ignored SNPs that have minor allele frequencies (5% or 10%) [148]. The second problem regarding rare alleles ($< 5\%$), is that they have little influence on the population as a whole even if they have a strong effect on the phenotype and therefore GWA has little power to detect them [151]. This issue might be particularly important in *A. thaliana* in which an excess of rare alleles has been observed [14].

Those later GWA drawbacks as well as genetic background confounding effects could also be partially solved by working at regional scales, which should decrease the variation in relatedness and the number of genes and of and the number of alleles at one loci contributing to the phenotypic variation observed for a given trait [148]. However, doing so will also increase average LD observed in the regional set [19] and so reduce the resolution of GWA mapping.

2.1.4 Nested association mapping (NAM)

As discussed before, GWA mapping has the advantage to have a high resolutive power but is sensitive to population structure. NAM was first developed in maize, where 25 founder lines chosen to maximize maize diversity and genotyped with 1.6 million SNPs were crossed with the B73 reference line to produce 25 RIL sets each composed of 200 individuals [152]. The resulting 5000 RILs were genotyped with common parent specific (CPS) markers (B73 rare allele) at relatively low density (1 marker /1.3 cM). With this population design, two different analyses can be run. First, joint linkage mapping analyses the phenotypic data of all RILs simultaneously in a single model including a term that accounts for the variation caused by the RIL family effect. Joint linkage mapping increases the power and the resolution of the mapping compared to the traditional linkage analysis. It also provides an estimate of the total variance associated with a region. In the second analysis, the sequencing data available in the parental lines are projected to the offspring within CPS marker intervals, which allows the use of historical recombination to improve mapping resolution. SNPs associations are tested across all the RILs in a joint analysis that also take into account the RIL family effect and the effect of the QTLs identified by joint linkage mapping. Overall, the combination of joint linkage mapping (CPS markers) and association mapping (parental SNPs) using the NAM population allows the direct identification of genes and/or SNPs contributing to a phenotype.

Besides, using RIL families permits the control of population structure, limits the number of false positives and negatives and allows repeated measures of phenotypes on the same lines, in common and different environments. However, the moderate number of founders limits genetic diversity and ancestral recombinations [153, 154, 155].

In *A. thaliana*, ampRILs and MAGIC lines have been created to bring more genetic diversity compared to standard RILs sets and now that 1001 genomes sequencing data are becoming available, combining GWA and linkage analysis in those populations will be possible [111, 112]. Besides, two NAM populations (figure 13) could be developed in *A. thaliana*. Two sets of 23 and 9 RILs populations obtained by the crossing of different accessions with Col-0 or Ler-0 common parents respectively are currently available (VAST lab) [107, 156]. Twenty one (5) out of the 23 (9) founders are currently being sequenced, which will allow GWA analyses but the RILs populations will also have to be genotyped with CPS markers to perform joint linkage analysis (several of the crosses to Col are currently being genotyped by sequencing). The combination of linkage and GWA mapping in *A. thaliana* has already shown to outperform each method used in isolation, linkage mapping being used to control the number of false positives and false negatives associated with GWA mapping and confirm specific segregation [141].

2.1.5 Validation of causal genes and polymorphisms

Once few candidate genes have been identified by mapping and cloning strategies, their specific effects on the phenotype have to be tested and validated. First, the phenotypic consequences of null alleles for the genes of interest can be investigated in the Col-0 background using the numerous T-DNA lines available in stock centers (see 1). Besides, gene silencing by RNA interference (RNAi) or artificial microRNAs (amiRNAs) is another convenient tool because it allows testing the activity of genes in other genetic backgrounds. However depending on the promoter of the construct used to introduce the RNAi or amiRNA into a given genotype or on the position of the transgene insertion in the genome, gene silencing efficiency may vary in different transformants and the expression levels of the genes of interest have to be tested to ensure silencing [157].

If altering the function of one gene modifies the phenotype of interest, quantitative complementation and knockdown approaches can be used to check whether the QTL is allelic to this candidate gene or not. The first method phenotypically compares the two QTL alleles against the TDNA allele in a heterozygous context (F1). If only one of the alleles complements the mutant phenotype (significant interaction between the alleles at the gene [WT/T-DNA] and at the QTL; figure 15) then the candidate gene is potentially responsible for the QTL [158], or at least interacting with it. Quantitative knockdown tests whether the inactivation of one allele by gene silencing differentially affects the phenotype compared to the inactivation of the second allele (significant interaction between the genetic background and the presence of the

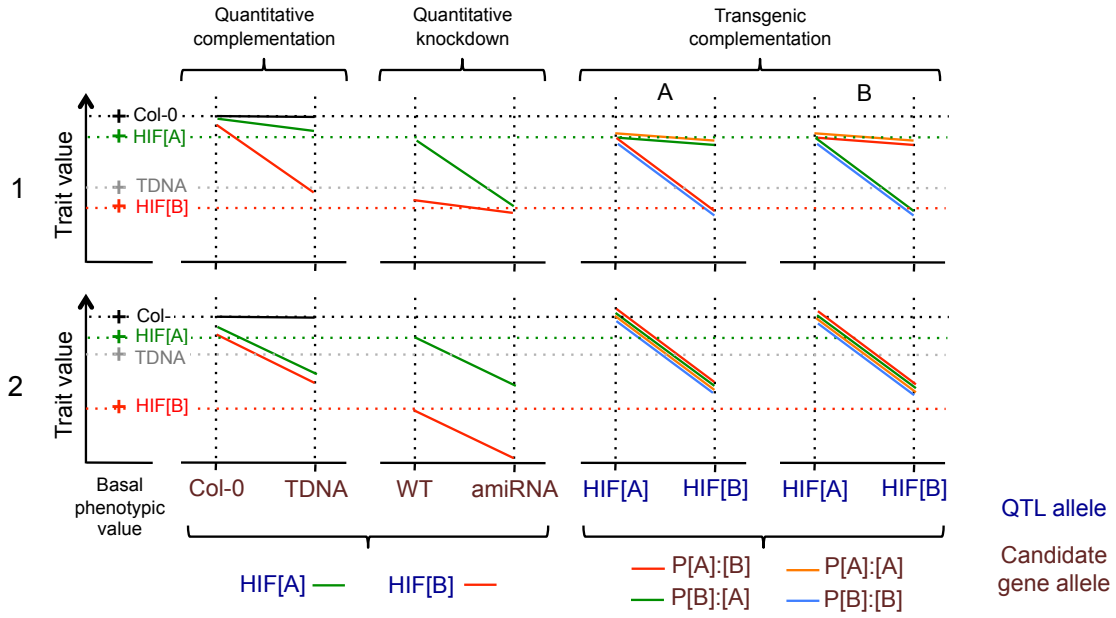


Fig. 15. Validation of a candidate quantitative trait gene. Representation of the phenotypic values expected in F1 after crossing an heterozygous HIF with an heterozygous T-DNA mutant of the candidate gene in Col-0 background (quantitative complementation), in fixed T3 after transformation of the HIFs with an amiRNA construction targeting the candidate gene (quantitative knockdown) or in fixed T3 after the transformation of the HIFs with the different alleles of the candidate gene under the control of native promoters (transgenic complementation). In this theoretical example, a recessive mutation in the HIF[B] and a recessive T-DNA insertion are responsible for a decrease value of the phenotypic trait. Overall, the candidate gene tested is responsible for the phenotypic difference observed between HIF[A] and HIF[B] if a significant interaction is observed between the QTL allele and the candidate gene allele (case 1). Using transgenic complementation, one can distinguish if the quantitative trait polymorphism is in the promoter (B) or coding (A) regions of the candidate gene. In case 2, the candidate gene tested is not responsible for the QTL. More complicated scenarios are possible.

transgene) (figure 15). If so then the candidate gene is probably responsible for the QTL [159]. Depending on the architecture of the trait, quantitative complementation and knockdown can be performed on the accessions directly or on the NILs/HIFs to avoid genetic background effects. As an alternative, transgenic complementation can help in the identification of the gene affecting a trait by introducing in different backgrounds a copy of different alleles and testing their effects on the phenotype (figure 15). Because the modifications associated to phenotypic differences can affect either protein functionality such as non-synonymous changes in functionally important amino-acids [160], frame-shifts [135], premature stop codons [161] and splicing alterations [162] or transcript accumulation due to DNA sequence variation [146, 163] or epigenetic modifications [62], the best is to introduce both the promoters and the genomic coding region of the genes of interest (figure 15). By introducing WT allele with point mutations, this technique can allow the identification and validation of the QTP, at least when it represents a single SNP. Despite the analysis of multiple independent transformants, the variation observed between different transgenic lines (as the result of insertions' position) may be larger than the QTL effect. Overall, the power of these three approaches strongly depends on the genetic architecture of the trait of interest.

2.2 The complex genetic architecture of quantitative traits

The genetic and recombination patterns characterising mapping populations and accession sets partially explain the statistical power of linkage and GWA approaches to detect genotype-phenotype associations. The other important factor is the genetic architecture of phenotypic variation, i.e. the number, identity, frequency and effect of all the polymorphisms affecting genes involved in the developmental, physiological and/or biochemical pathways associated with a trait phenotype, as well as all the 2 ways and higher-order epistatic interactions. The more complex the genetic architecture underlying the variation observed in a trait is, the more difficult it is to study it and to identify all its components. In this section I will show that linkage and GWA mapping differ in their ability to identify some of those components (figure 16). However because they are inherent to quantitative traits, it is important to be able to develop statistical methods in order to better detect them.

2.2.1 The number and effects of QTLs

Thanks to linkage mapping and GWA analysis it is possible to identify the QTLs that compose the heritability of a trait. In theory, the sum of the % of phenotypic variation explained by all QTLs should be equal to heritability but it is rarely the case. This difference is partially due to epistasis–on which we will come back later–but also to small-effect QTLs that are difficult to identify individually but globally contribute to explain a given fraction of phenotypic

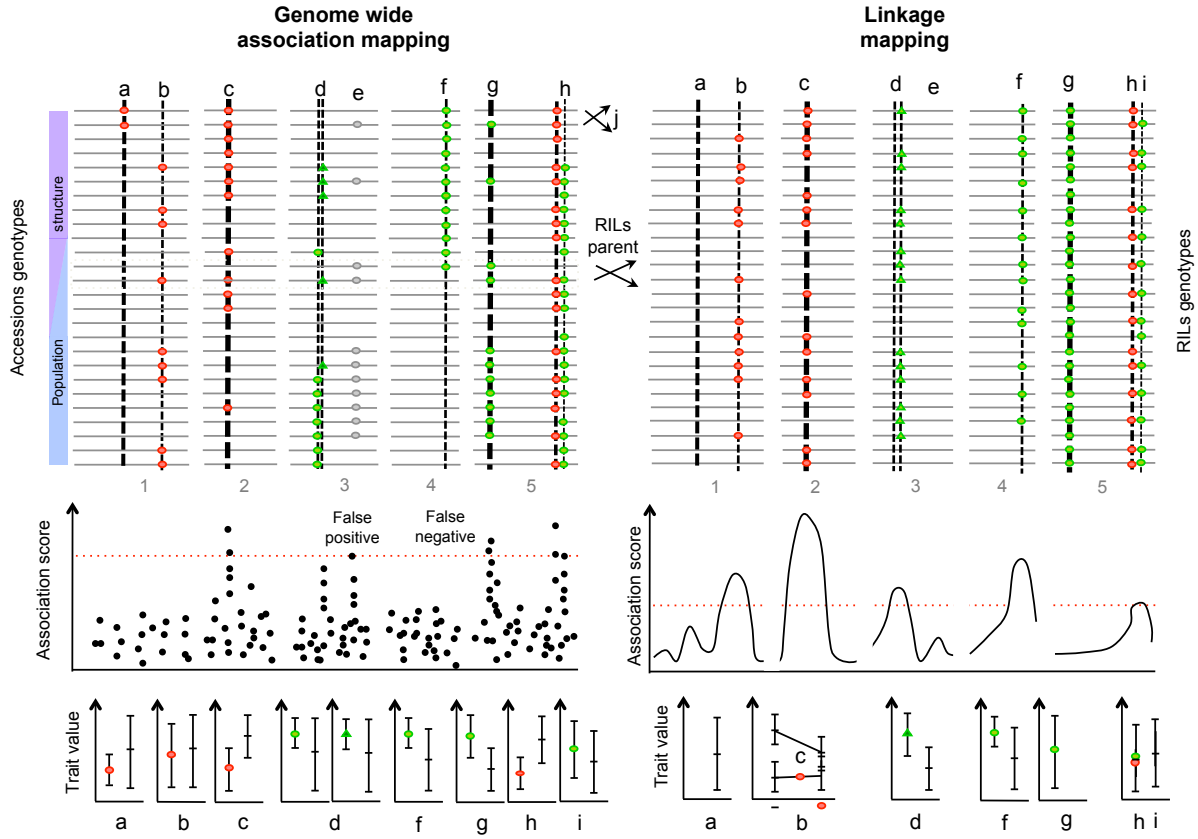


Fig. 16. The effect of genetic architecture on the power to detect significant associations using linkage and GWA mapping. The genetic architecture of a theoretical trait is represented in a set of accessions and in a RILs population. On the five chromosomes of each individuals (grey horizontal bars), the alleles with positive (green) or negative (red) effect on the trait are indicated on vertical dashed bars (QTL position) whose thickness is representative of the strength of QTL effect. The association score associated with both analysis and the phenotypic trait values associated with each QTL alleles are also indicated. a – Association mapping have poor power to detect rare variants that will be detected using linkage mapping only if the variant segregate in the parents of the RILs population. b-c – Association mapping may not detect some loci because they are in epistatic interaction. Epistatic loci can be detected using special statistical analysis and linkage mapping data when the population is big enough or the interaction very strong. d – Allelic heterogeneity reduce the power of association mapping. e-f – The genetic background effect can cause spurious associations (false positive). Here polymorphism e is in linkage disequilibrium with polymorphism g and so is coming out in the GWA analysis. Besides, after correction for population structure, some loci may not be detected anymore (false negative) because of correlations between QTL allele distribution and population structure (f). g – Depending on the alleles segregating in the parents of RILs population, loci detected using GWA mapping may or not be detected by linkage mapping. For example in the RILs population issued from the cross j, only one QTL (g) would have been detected using linkage mapping. h-i – Linkage mapping has a poor power to detect linked QTLs compared to GWA, more particularly when the they have opposite effects.

variation. Regarding that, if the variation in a trait is mainly explained by many QTLs of low effect, they will almost be undetectable. In *A. thaliana*, the observed distribution of QTLs effects is mostly L-shaped with few loci of large-effect and many small-effect QTLs, but varies depending on phenotypic traits. For example pathogen resistance is usually governed by major dominant loci [74, 164] whereas mineral concentrations are usually controlled by more QTLs of relatively small effects [74, 97, 165, 166]. Besides, the L-shaped distribution of QTL effects could result from biases due to small sample size (the Beavis effect), linkage (see 2.2.3) or epistatic interactions (see 2.2.4) [158].

Because accession collections capture more genetic variants than mapping population, the architecture of the traits in those sets is likely more complicated with more QTLs likely explaining a smaller fraction of the total phenotypic variance (also the total phenotypic variance is likely higher in accession sets) and less observed replicate per QTL allele than in a simple RIL set. As a consequence, small effect QTLs are likely more difficult to identify in GWA. Besides, most of the current models used in GWA analysis, test the effect of each SNP individually without taking into account the other QTLs, which increases the background variance associated to small effect QTLs due to large effect QTLs. This may partially explains why GWA mapping was more performant to identify genotype-phenotype associations for resistance to pathogens than for ionomic traits [74]. A new model has been developed to better take into account loci of larger effects and a priori knowledge from *A. thaliana* linkage mapping analyses [167]. This model identifies new associations as well as evidence for allelic heterogenetity (see 2.2.2).

2.2.2 Rare variants & allelic heterogeneity

We discussed earlier that rare variants are poorly detected by GWA analyses. Because they are rare at the worldwide species scale, those variants are less likely to be found in mapping populations. However, if they are, they will be identified by linkage analysis. Several rare variants such as *RAS1[Sha]* and *APR2[Sha]* have been identified as QTLs thanks to linkage mapping [161, 160]. The impact of rare alleles at the species scale could appear less important compared to frequent alleles. Nevertheless, any variants positively contributing to fitness could in fine be fixed in a population and so from a predictive point of view, one might be interested to discover all the polymophisms seregating in a population and that could contribute to the species evolution. Beyond this genetic heterogeneity, it is possible that low frequency mutations appeared independently in the same gene in different accessions and that they all contribute significantly to phenotypic variation at the species scale. Such allelic heterogeneity [9] has been observed for several genes such as the major flowering time regulators *FRI* [74, 168, 169, 170, 171] and *FLC* [60, 169, 170, 171, 172, 173, 174] that independently accumulated several non-

9. Allelic heterogenetity is observed when different mutations in the same gene lead to similar phenotypic variation

synonymous mutations, premature stop-codons, deletions, insertions and non-coding mutations; the *HMA5* encoding a Cu transporter responsible for Cu tolerance [175] and *HMA3* a major locus responsible for leaf Cd accumulation [176].

Allelic heterogeneity is not a major problem for linkage mapping analyses (at least in 2-way crosses) but to be detected the mapping will have to be performed in at least two crosses segregating for the different causal variants. Alternatively, allelic heterogeneity can be detected a posteriori when analysing haplotypic diversity of a candidate gene in accession sets. Regarding GWA, allelic heterogeneity is likely to reduce the power to detect significant associations by increasing the phenotypic variance at causal alleles (and also potentially by decreasing the frequency of each causal variant in the set). Nevertheless significant associations have been observed for *FRI*, *FLC* [74] and *HMA3* [176] indicating that this trend is not a rule and depends on traits architecture. Maybe more importantly, when two different alleles in a gene contribute to a phenotype, a SNP common to the two alleles and segregating in a close gene could turn out to be better associated to the phenotype than the two actual causal variants and could lead to misleading conclusions regarding the candidate gene (Nordborg, personal communication). But recently-developed algorithms could partially solve this problem [167].

2.2.3 Pleiotropy or linkage

In part 1, we saw that phenotypic traits are not completely independent one another. Within populations, these correlations may reflect common genetic bases which should then be supported by QTL colocalizations between traits. Several QTL colocalizations have been identified within and between levels of biological organisation for mineral concentrations [92, 97, 166], primary and secondary metabolites contents [132, 133, 178, 179], protein abundance and activity as well as for growth-related traits [177]. The most straightforward explanation for QTLs colocalization is the presence of a pleiotropic locus, reflecting true independent effects of one locus on distinct traits or causality relationship between traits after one is affected by the locus. For expression and metabolic QTL several regulatory hotspots have been identified likely corresponding to important pleiotropic loci [179, 180, 181] and reflecting the 'hub and spoke' nature of the underlying networks [179, 180]. A major pleiotropic locus observed mostly in mapping populations involving Ler and Van-0 accessions is the *ERECTA* gene which encodes leucine-rich repeat receptor-like Ser/Thr kinase. This gene has been identified as a major trans-regulator of several genes involved in biotic and abiotic response [180] and also control various metabolite and protein QTLs [177] as well as epidermal cell expansion and division and leaf growth [182].

The absence of QTL colocalization is not a proof per se that the two traits do not share common genetic bases. Indeed, antagonistic QTLs, which are QTLs with opposite effects on pair traits, could mask the effect of other pleiotropic loci [183]. Conversely, QTLs colocalization

could also be due to genetic linkage between two different loci with independent phenotypic effects. For example, biosynthetic genes sometimes appear in clusters due to gene duplication, which results in a QTL hotspot for metabolites variation [184]. Whereas linkage issues are expected to be more important when QTL are mapped using linkage mapping than GWA mapping (because accessions accumulated more recombinations than RILs populations), this is actually not always the case. For instance, because of the strong LD observed in accessions sets around *AOP2/AOP3* and *MAM1/MAM3* genes, those loci appeared as hotspots of associations with GSL traits [185]. Mixed models can be used to take into account the data from multiple traits in QTL mapping analyses [186] and new models are under development to use the data of several correlated traits to improve detection power in GWA analyses [187].

2.2.4 Epistatic interactions

There are two main definitions to the term epistasis. The first one, used by molecular geneticists and introduced by William Bateson in the early 1900s refers to the masking of an allelic effect by an allele at another locus in a given homogeneous genetic background [188]. This type of epistatic interaction has been particularly useful to find out if genes belong to common, additive or synergistic pathways and in the former case to organise genes' effects within biochemical or regulatory pathways. The second definition introduced by Fisher and used by population and quantitative geneticists refers to any statistical interaction between the genotype at two or more loci within a population [188]. This statistical deviation from additivity is dependent on all the other loci segregating in the population that may confound the relative effects of the two interacting loci. This is why not detecting epistasis in the statistical sense does not mean that there is no interesting interaction in the strict genetic sense. Similarly, finding epistasis in one genetic context does not mean that there will be epistasis in other contexts (statistically).

In linkage QTL mapping studies, epistasis started to be analysed thanks to the development of specific statistical methods such as 2D scans and MQM analyses (see above). However, several limitations to the detection of epistasis remain. First, to test for pairwise gene-gene interations, the mapping population has to be splitted in 4 genotypic classes instead of 2 for main effects QTLs, which results in fewer replicated observations per class and, hence, less power. Then, testing all possible pairs of markers is time consuming and implies correcting for multiple testing which strongly raises the significance threshold so that only extremely strong interactions can be detected. If the two epistatic QTLs are linked, QTL mapping will have a very low power to detect them but using series of NILs can help identifying linked epistatic QTLs [189]. Finally, the segregation of other QTLs can interfere with the detection of epistasis between the pairs of markers under consideration. Considering the difficulty of mapping epistasis in controlled crosses that strongly diminish the complexity of the genetic architecture of phenotypic traits,

epistatic interations are simply not taken into account in GWA mapping. Considering all the pairwise comparisons that would have to be tested, special algorithms need to be developed and the number of observations needed is likely to be overwhelming.

Overall, numerous epistatic interactions have been observed for various traits in *A. thaliana* mapping populations [190, 191, 192, 193, 194, 195, 196] suggesting that epistasis play a major role in plant genetic architecture. However, apart from the duplicated genes or immune responsive genes involved in post-zygotic incompatibities between *A. thaliana* accessions [62, 197, 198, 199], few gene-gene interactions have been mapped down to the gene and understood at the molecular level. One of the reasons for that is suggested by the two definitions presented earlier. To confirm and fine-map an epistatic interaction, one as to pass from a heterogeneous background (mapping population) to a homogeneous one (HIFs or NILs) but the studied epistasis might be dependent on the genetic context of the mapping population [132]. Similarly, epistatic interactions identified in mapping populations might not be relevant at the species scale [200].

If mutational studies are not interesting per se to identify the residual variation associated to with epistatic interactions in natural populations, they provide interesting biochemical and regulatory frameworks to understand epistatic interactions. For example, Rowe et al identified several recurrent gene-gene interactions with positive or negative effects for the accumulation of different metabolites, which allowed them to highlight where in specific pathways were located the SNPs responsible for the interaction [179].

Finally, higher orders of interaction, such as three-way epistasis have not been assessed genome-wide because the number of combinations that have to be tested is too high, and would take much more individuals than is currently typically used. Nevertheless, it is possible to test the interaction between specific loci in MQM models or simply by ANOVA. Once again, despite the limited power, those interactions are often identified in *A. thaliana* [179] but more analyses need to be performed to better learn how much they contribute to phenotypic variance.

2.3 The unexpected complexity of phenotypic variation.

2.3.1 Phenotypic buffering

Integrating the information obtained from QTL analyses at different levels of biological organisation can help understanding how genomic, epigenomic, transcriptomic, proteomic, metabolomic and developmental pathways interconnect each others. Several studies have combined in a row the analysis of transcriptome and metabolome variation in RILs populations [178, 179, 201] and one analysis also included proteomic and morphological traits [177]. For example, Keurentjes and colleagues showed that variation in metabolites accumulation could be explained by variation in enzymatic activities due to genetic mutations in coding sequences

or variation in enzyme quantities due to variation in transcript accumulations [178]. Overall, the simultaneous analysis of different levels of organisation can help reconstructing functional networks and reveal the impact of genomic mutations across different phenotypic levels, as it has been shown for example for glucosinolate [202]. Besides, Fu and colleagues showed that most of the huge genomic (likely including epigenetic) diversity observed between Col-0 and Cvi-0 accessions is subject to a strong genetic buffering at higher order of biological organisation. Indeed, whereas many expression QTLs have been isolated in Cvi-0 population only few of them have been shown to propagate to higher phenotypic levels, the candidate genes of those rare hotspots being central cellular network hubs that control most of the phenotypic variation observed [177]. Besides, heritability values of metabolic traits are lower than the one of transcriptomic traits, which might reflect the higher complexity of phenotypic traits as one moves away from genotype to phenotype [179]. Phenotypic robustness is an important aspect of biological systems as it ensures the proper development of organisms despite mutation accumulation or hybridization. This cryptic variation buffered in a given organism can turn out to be essential in other genetic and physical environments.

2.3.2 The genetic x environment interactions

As mentioned earlier, plants developed different responses to different environmental constraints and one important question is whether variation in trait plasticity is driven by specific or shared QTLs across environment. Because *A. thaliana* accessions and mapping populations are rather homogeneous, it is possible to phenotype them repeatedly and so to perform QTL mapping analyses in various environmental conditions. Statistically, the search for QTLs can be performed independently in several environments using the methods described above and models compared afterwards. Conversely, mixed models can be used and directly integrate in a single model QTL, environment and QTL x environment effects [186].

Often, many of the isolated QTLs are specific to a given condition. For example, the comparison of the eQTLs identified in Ler^{erecta} x Cvi-0 RILs population on standard conditions or after 3 hours of shade revealed that 91% of the eQTLs identified in both experiments were specific to the environment, distant/trans eQTLs being the most affected [203] although there could be some power issues here. Similarly, Chan and colleagues showed that glucosinolate-related metabolites QTLs could be specific to the environment under study as well as the organ considered [204] and the same conclusions were reached regarding ionomic traits [92, 97, 165, 166]. Finally, growth and flowering QTLs are also sensitive to environmental variation even though QTLs pleiotropic across environments have been isolated (*CRY2*, *FRI* ...) [141, 142, 205, 206]. Even when QTLs are detected in several environments, genotype x environment interaction can exist as a result of change in the magnitude of the QTL effect rather than change in rank order effect. For example, under growth-limiting nitrogen conditions the *SO3.1*

locus increased its impact on sulphate content, indicating that the strength of the QTL control on a trait–or pathway–can be environment-dependent [160].

Overall, these observations likely reflect different types of loci involved in plasticity. First some genes directly influence phenotypic traits whatever the environment but their importance might vary in certain conditions as a result of differential regulation (pre or post transcriptional). In that case, selection of the trait will altered plasticity and vice versa. Conversely, some proteins/genes might specifically affect plasticity in a given trait in response to an environmental cue so that selection for the trait won't affect plasticity and vice versa [208].

Because phenotypic integration likely constraints phenotypic plasticity as a result of trade off between traits, identifying the network of characters correlations within and between levels of biological organisation will be essential to predict the evolution of plasticity in changing environments [207]. Besides, it could strongly help identifying how a molecular variant globally affects the pattern of phenotypic variation in different genetic backgrounds. Nevertheless, the environment may also modify correlations among traits [208] and so identifying these networks might be even more complicated than presented before [205].

2.4 Conclusion about QTLs detection using natural variation.

Depending on the method used and traits' genetic architecture, QTL analyses allow the identification of the loci, genes and in the best cases polymorphisms responsible for the natural phenotypic variation observed in various environmental conditions (experimental or natural).

One of the questions that are often asked is what natural variation tells us about complex traits compared to mutant analyses. The first argument is that most mutants are available in a single genetic background (most of the time Col-0 or Ws-0) so that the phenotypic variation observed in mutant populations does not account for epistatic interactions that can increase or decrease and often cancel the effect of some loci (exemple: gene redundancy). Then, mutations with quantitative effects and that are sensitive to environmental variation can only be detected through the phenotyping of multiple individuals bearing the same mutation. Mutant screens for variation in a quantitative trait would then be possible with T-DNA mutants for which the position of the insertion is identified independently of its phenotypic effect but not for ems populations. But, in any case, the phenotyping work in mutants is much less efficient than with RILs or accessions, since a single genetic event (mutant) requires phenotyping one line, while each accession or RIL observed allow to test multiple mutations (in combination). In addition, one problem with T-DNA mutant is that they are often knock down for the gene of interest whereas accessions accumulated several mutations with relatively small effects which may increase phenotypic diversity and reveal the network of characters correlations. In fine, analysing natural variation can lead to the discovery of new regulatory pathways such as the

association between the xylem expansion and flowering induction [209]. Finally, regarding evolution and the possible adaptive potential of genes, nothing really makes sense except in the light of natural populations and environment!

3. EVOLUTIONARY SIGNIFICANCE OF *A. THALIANA* NATURAL VARIATION

In the previous chapter, we showed that linkage mapping and GWA analyses can give us precious information about how genetic, phenotypic and environmental variables interact and result in the observed complex traits in natural populations. A next step is to understand the ecologically and evolutionary significance of this complexity. This includes identifying the environmental variables that affect fitness-related traits and the genetic variants that are selected in consequence. Here selection refers to the process that will increase the frequency of genetic variants (and so of the associated phenotypes) through generations at the worldwide and species level or at more local scales in a given environmental context. As mentioned in the first part, *A. thaliana* is likely to have a huge capacity of adaptation underlined by its large geographic range and its occurrence in very diverse habitats. As a consequence part of the genetic and epigenetic diversity observed in that species and the resulting phenotypic variation is likely to be adaptive (figure 17). In this chapter, I will present how we can identify the natural phenotypic variation that matters regarding fitness and natural environments and so that is likely to be adaptive. Then, I will show that selection processes can profoundly affect *A. thaliana* genetic diversity, which can then be used to detect traces of adaptation at different species and time scales.

3.1 Ecologically significant phenotypic variation in A.thaliana.

Fitness is the group of traits on which selection acts and so fitness differences between genetically different individuals in one or several environments are essential for adaptation to occur. Fitness (noted W) refers to the ability of an individual to survive and reproduce in a given environment and so to contribute to the genetic pool observed in the population at the next generation [210]. During plants life cycle, after germination, seedlings have to survive to various biotic and abiotic constraints until the production of the first fruits (survival) and then have to produce a descent number of viable seeds (reproduction) that will follow the same cycle. Because fitness results from many different components that can be difficult to identify, fitness is tough to measure and most of the time is only estimated from phenotypic traits associated to survival or reproduction. Some estimates are more intuitive than others. For example, in *A.*

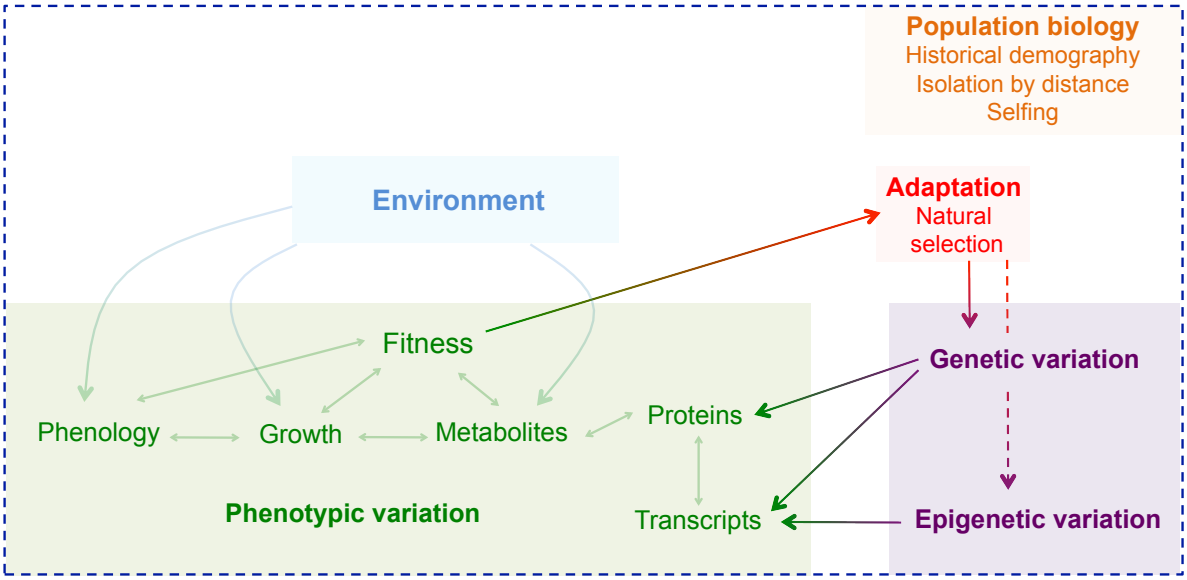


Fig. 17. Plants adaptation to given environments through the selection of advantageous genetic variants. Illustration of the topics aborded in the chapter 3 of the introduction.

thaliana, lifetime fitness is often estimated from the total number of fruits (siliques) produced per plants and this estimates often takes into account survival by assigning the value 0 to the individuals that do not survive [211, 212]. However, this value does not consider the variation in the number of seeds produced by each silique in genotypically different individuals nor the viability of the seeds produced.

Quantitative geneticists are often interested to know if the trait they are analysing is a component of fitness and so if its variation is associated to adaptation or not. The standard approach consists in analysing correlations between phenotypic traits and fitness in a genetically heterogeneous population. But special care must be taken when inferring causal relationships from those correlations because phenotypic traits differences may covary with fitness due to developmental or incidental reasons but not give rise to fitness differences [213]. QTL colocalization can help identifying the phenotypic traits that underlie fitness variation. For example, QTLs colocalizations for total life time fitness and seed germination phenology traits suggested that primary seed dormancy might be critical in the early process of adaptation in *A. thaliana* [212]. This result makes sense regarding the importance of germination timing (spring/fall) for the survival of young seedlings and for defining the seasonal environment experienced by the rosettes at all subsequent stages [214] that will strongly influence life history traits [215, 216]. However, here again, QTL colocalization is not an absolute proof of the causal link between phenotypic traits and fitness because pleiotropic QTLs (if not linked QTLs) could influence two characters independently [213]. Besides, if two genetically correlated traits have different effects on fitness, their evolutionary strength will be constrained. Overall, it is the sum of all

the loci affecting fitness and their interactions that will determine the response of plants to evolutionary forces.

Like many phenotypic traits, fitness fluctuates through time and space as a result of changing environmental conditions [210]. As a consequence identifying a variation in fitness and the underlying QTLs may hold only for the environment under study. This is particularly important regarding the fact that most QTLs cloned so far have been identified in artificial conditions and it is clear from several experiments that the QTLs detected under controlled laboratory conditions can differ from the one identified in field experiments [141, 192, 217]. The best experimental design to follow regarding QTL mapping experiments is unclear. On the one hand, analysing phenotypic variation in natural environments is more likely to reveal the QTLs that do matter in nature but several hidden environmental variables are expected to increase the phenotypic variation due to non genetic factors which may in turn decrease detection power. Besides, because several environmental parameters will be acting at the same time, we are not sure to identify the QTLs that are related to the environmental variable we want to study. On the other hand, identifying QTLs in controlled conditions should reduce phenotypic variation due to hidden environmental factors and increase power but is likely to detect QTLs that do not necessarily make sense in terms of natural environments. An intermediate solution could be to grow plants in seminatural environments such as in growth chambers that reproduce light and temperature variations observed in natural climate [142, 206] or outdoors but under controlled and uniform soil conditions [141, 211]. However those solutions remain quite disputable as small variations in growing conditions, even with standardize protocols and under controlled conditions, can account significantly to the observed phenotypic variation [218].

The environment does not only strongly affects phenotypic variation. It is the selective agent that causes natural selection and adaptation. Thus, correlations between phenotypes or ultimately QTLs allele frequencies and environmental components can be a good indication of adaptation. In *A. thaliana*, such correlations have been observed for several light sensitivity and flowering time traits and underlie QTLs that show latitudinal, longitudinal and altitudinal clines [173, 219, 220, 221]. However, special care must be taken when analysing this kind of topoclines because population structure could partially explain such correlations. To discard this possibility, Balasubramian and colleagues showed that *PHYC* which occurs in two natural haplotype groups varying for flowering time, displayed a greater latitudinal cline variation compared to others SNPs distributed throughout the genome [222]. Alternatively, a mixed linear model that included genetic relatedness was used to test the association of 35 flowering QTP and various environmental variables [173]. The use of the latter mixed model clearly showed that false-positive associations could arise from population structure. Variations in the detection of flowering time topoclines in experiments using different *A. thaliana* accession sets and different environmental settings suggest that the power to detect clines is strongly influenced by

the sample composition (and underlying alleles frequencies) and experimental environmental conditions [220, 221]. For example, a latitudinal cline for flowering time among 70 european accessions when grown under ecologically realistic conditions was shown to be dependent on the functionality of *FRI* alleles [223]. To better understand which climatic factors drive topoclines and so the evolution and distribution of flowering time variation, the precise description of the annual and seasonal temperatures, precipitations, day length, cloud cover, light quality... recorded over the last decades have been precious (WorldClim, CRU/UEA). The association of those bioclimatic variables with phenotypes or genotypes using multivariate analyses or specific modelling approaches does not only allow identifying the climatic factors likely acting on the selection of particular variants [173, 224]. They also determine the most suitable habitats (ecologically and geographically speaking) associated with important genetic variants, which could in turn be used to predict how genotypes will respond to changing environments [225, 211]. Although clines associating phenological traits (germination and flowering) with temperature and precipitation climatic factors start to be revealed [226, 73, 173, 224, 225], the clines associating survival and edaphic or biotic parameters are poorly understood mainly because of the geographic heterogeneity of those parameters. Baxter and colleagues observed a significant enrichment of a weak allele of *HKT1;1* – encoding a sodium transporter – in region associated with coastal and saline soils in Europe [145]. However the exact information about soil salinity at the place of origin of the accessions used in that study was not available and so conclusion on the adaptive value of *HKT1;1* weak allele might be elusive. Overall, there is very few information about soil composition and pathogens present where accessions have been collected and efforts need to be done to better describe these local ecological variables in order to understand plant adaptation to biotic and edaphic constraints [227, 228].

Artificial selection experiments could give some insights about the selection of alleles associated with increased fitness in particular biotic and edaphic conditions [229, 230, 231, 232]. An alternative method could be to perform the classical reciprocal transplants between several natural environments known to vary or not for the environmental variable of interest. This approach is a gold standard in ecology and is invariably successful in *A. thaliana* to reveal local adaptation [211, 214, 217, 233, 234, 235]. For example, thousands of accessions were phenotyped for fitness-related traits in 4 common gardens located across *A. thaliana* species' native range. Compared to genomic controls, high fitness alleles – detected by GWA analysis – were generally distributed closer to the site where they increased fitness providing indirect evidence for local adaptation [211]. Similar results were obtained from the reciprocal transplant of two European accessions from Sweden and Italy showing that the fitness of the local population is globally higher than the one of the non-local accession [235]. Those patterns of local adaptation could be explained by a combination of two genetic mechanisms. First, **antagonistic pleiotropy** occurs when one QTL allele is associated with increased fitness in one environment

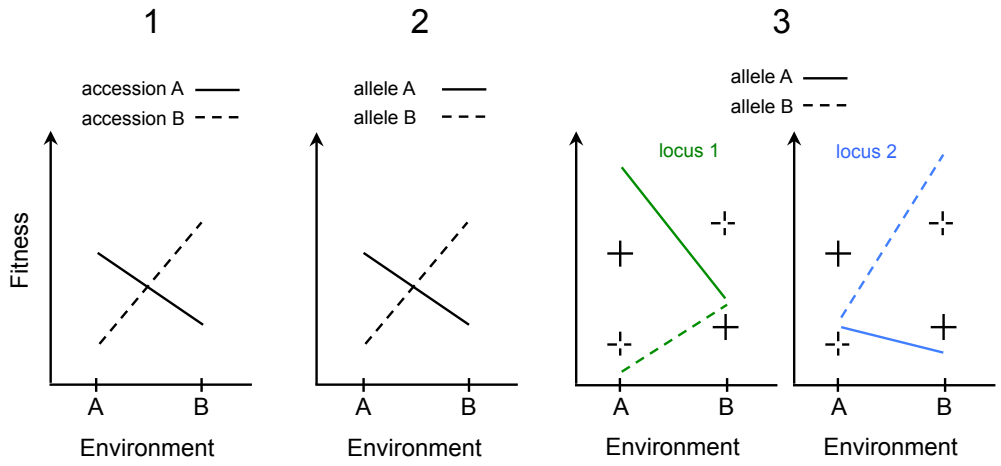


Fig. 18. Local adaptation in accessions (1) can be explained by antagonistic pleiotropy (2) or conditional neutrality at two different loci (3). Each plot represents the fitness norm observed in two different environments. Accessions A isolated from location A characterized by a given environment (A) has a higher fitness on its native environment compared to accession B and vice versa. This trade off can be explained by two alleles of one QTL being advantageous in one environment but deleterious in the other (2) or two QTLs with one allele advantageous in one environment and neutral in the other one (3). Such reaction norms (2 & 3) could be obtained by the phenotyping in natural environments of near isogenic lines fixed differently for one or two loci.

compared to the other QTL allele ($W_A > W_a$) but that the opposite is observed in another environment ($W_A < W_a$) (figure 18). The evidence of alleles, associated with high fitness in a given site but having a restricted geographic range likely due to negative pleiotropic effects in other environments, provides good evidence of local adaptation [211]. However changes in the rank of fitness is not necessary for local adaptation to occur. Indeed, **conditional neutrality** which is observed when the magnitude but not the rank of the QTL alleles change between different environments, could also explain patterns of local adaptation when acting on several loci (figure 18). Conditional neutrality is observed more frequently than antagonistic pleiotropy [205, 211, 236]. But for now it is not clear if this results from a poor experimental and statistical power to detect antagonistic pleiotropy or if prevalence of conditional neutrality makes sense on an evolutionary perspective [236]. Interestingly, the genetic bases of fitness vary considerably across sites suggesting that local adaptation act on different target loci and different molecular processes in different environments [211]. Change in fitness across environments may be the result of environmental plasticity and plasticity costs. But plasticity is not always selected for and selection for plasticity may vary depending on the strength and duration of environmental perturbations [207]. Finally, it is important to note that absolute fitness differences may vary through time because of the evolution of selective pressures [235]. It has been suggested that natural selection favours alleles with a smaller variance in fitness through time [210] but it re-

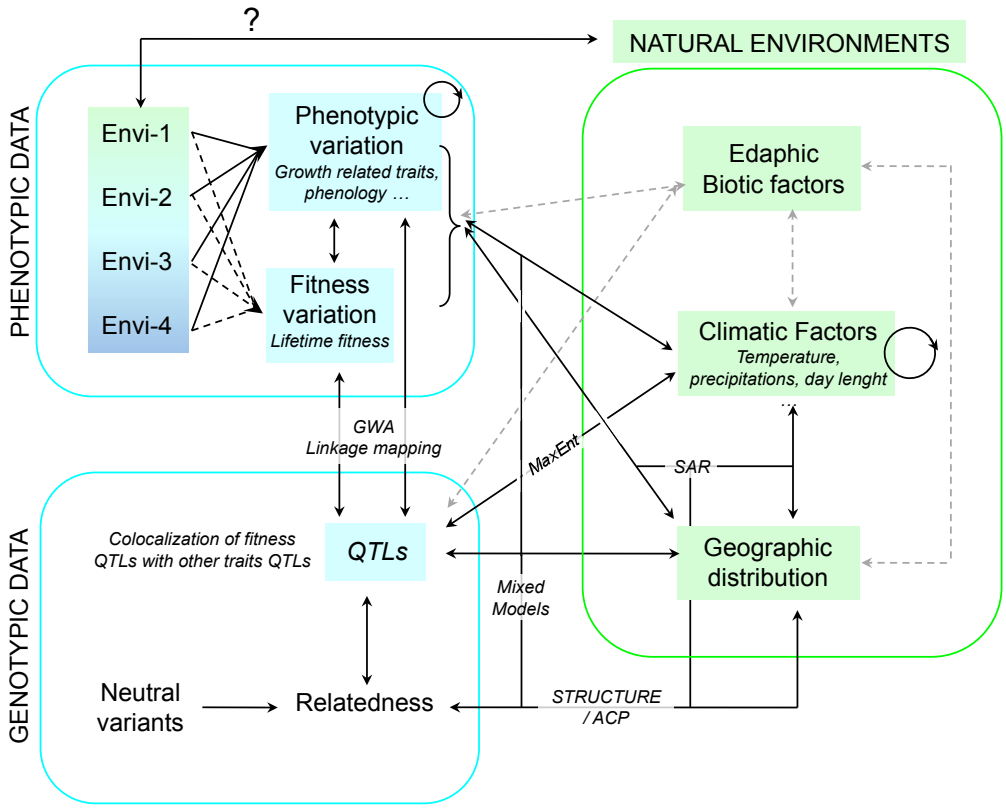


Fig. 19. The analysis of local adaptation generally requires the identification of loci associated with fitness variation in multiple experimental environments using GWA or linkage mapping in several populations. Then the distribution pattern of adaptive variants i.e. their environmental and geographic specificities have to be analysed and confronted to known pattern of population structure (relatedness) observed in *A. thaliana* in order to identify the selective agents that drive local adaptation. Some of the software that can be used to analyse correlations between the phenotypic, genetic and environmental variations are indicated. The dashed grey arrows represent correlations that have been poorly analysed. Note that one of the difficulty is to know whether the phenotypic variation and underlying QTLs that are observed in experimental environments are similar to the one observed in natural environments (question mark, blue vs green).

mains to be determined if this rule also prevails in *A. thaliana* populations. The phenotyping of naturally growing accessions in situ over several years will be essential to answer this question [11, 237].

As a conclusion, the analysis of local adaptation generally requires the identification of loci associated with fitness variation in multiple natural environments characterized by different climatic, biotic and edaphic parameters using GWA or linkage mapping in several populations (figure 19). Then the distribution pattern of adaptive variants i.e. their environmental and geographic specificities have to be analysed and confronted to known pattern of population structure observed in *A. thaliana* in order to identify the selective agents that drive local adaptation (figure 19). Finally, transplant experiments are recommended to confirm the observed

associations. Ultimately, global knowledge on the selective forces that drives adaptation could be used to predict the evolution of traits and genetic patterns under changing climates. An approach complementary to ecological methods is to directly search for loci under selection using population genetics.

3.2 Evolutionary significance of phenotypic variation

Basically, population genetics approaches use molecular data to identify signatures of natural selection by comparing the observed pattern of nucleotide diversity with the one expected under a neutral evolution hypothesis. The accumulation of genetic data (See table 1) has been important in the evolution of those methods that use genome-wide SNPs informations to infer selection. Genes or polymorphisms associated with phenotypic variation have long been tested individually for selection. This approach provided precious information notably regarding pathogen defence-associated QTLs for which ecological data was not necessarily available. Using whole genome sequencing information, it is now possible to perform genomic scans on a population with no prior information on the possible traits or genes under selection. In this subchapter, I will first present the neutral theory and different modes of selection underlying how they affect the patterns of genetic diversity. Then, I will go through different methods, the so called 'tests of selection' , that can be used to infer selection at the genome wide or gene levels. Several reviews have been used and can be consulted for more details [238, 239, 240, 241, 242]

3.2.1 Theoretical elements

The **neutral theory** first proposed by Kimura in 1968, stipulates that most of the genetic variants observed within and between species have no effect on fitness. The frequency of those selectively neutral variants is determined randomly by the stochastic effects of genetic drift and population biology. Associated with two other assumptions i.e. random mating and constant long-term population size, the neutral model can be used to predict the pattern of polymorphisms that would be observed if a locus is neutrally evolving. This reference pattern is often considered as the null hypothesis in selection tests and can be compared to observed patterns of polymorphisms. In selection tests, rejection of the neutral hypothesis can be interpreted either by selection or by demographic forces that would not have been considered by the neutral model [238].

In theory, different types of selection can be distinguished. First, **negative selection** (or purifying selection) [10] quickly acts on the removal of deleterious mutations (i.e mutations that affect negatively plant fitness ($W_{aa} < W_{Aa} < W_{AA}$, a being the derived allele)) and

10. Negative selection is observed when the frequency of a deleterious mutation, reducing fitness, decreases in the population

consequently reduce both inter- and - to a lesser extend- intra-specific variability [239]. The reduced rate of divergence can be used to search for conserved region within the genome that are likely functionally important. Besides, because deleterious mutations are more likely associated with amino acids replacements than with synonymous substitutions, negative selection leaves a widespread signal in the genome consisting of an enhanced rate of silent vs non-synonymous mutations. Overall, deleterious mutations are likely to contribute little to the observed levels of polymorphism under the neutral model. Nevertheless, there is a balance between mutation and selection processes that is influenced by population size and the magnitude of allelic effects on fitness. As a result, deleterious mutations can segregate transiently at a low frequency within populations before being removed by selection. This phenomenon is likely to be more important in small populations where selection is less efficient [240] and/or in species where the environmental constraints are likely to be constantly evolving, for example with pioneer species. In *A. thaliana*, several rare variants with negative effect on fitness in laboratory conditions have been isolated suggesting that in this selfing species, deleterious variants are partially contributing to phenotypic diversity [160, 161].

Directional positive selection[11], which is of much more interest for population biologists, occurs when a variant with positive effects on fitness ($W_{aa} > W_{aA} > W_{AA}$, a being the derived allele) is favoured by selection. Overall, positive selection results in an increase of the advantageous allele and linked variants frequency that can eventually get fixed at the species level or within a population facing the selective agent (local adaptation). Depending on the strength and scale of selection, the selected variant will be fixed more or less rapidly, and will drag along more or less linked variants (hitchhiking, see below) resulting in more or less reduced genetic diversity around the favoured site or **selective sweep** [238] (figure 20). When a new allele is fixed rapidly, one uses the term of **hard sweep**. Besides, because of their relatively young age, new loci that are undergoing or just undergone selection are expected to show a relatively high level of LD characterized by long haplotypes that have not (yet) been eroded by recombination [243]. Finally, genomic regions that recently swept present an excess of rare polymorphisms that correspond to recent mutations that occurred after the sweep. Sweeps do not always arise from new advantageous mutations. A neutral variant that has already been subjected to drift generating new haplotypes through recombination and mutation or an allele originating from several independent mutations can, under changing environment, become advantageous and sweep to fixation. In that case of **soft sweep**, the background signal will be similar but weaker than under hard sweeps. Overall, it is important to keep in mind that positive selection is an evolving process. As a result the pattern of polymorphisms of alleles undergoing positive selection (**partial sweep**) will be different from the one of mutations that

11. Positive selection is observed when the frequency of an advantageous mutation, increasing fitness, increases in the population

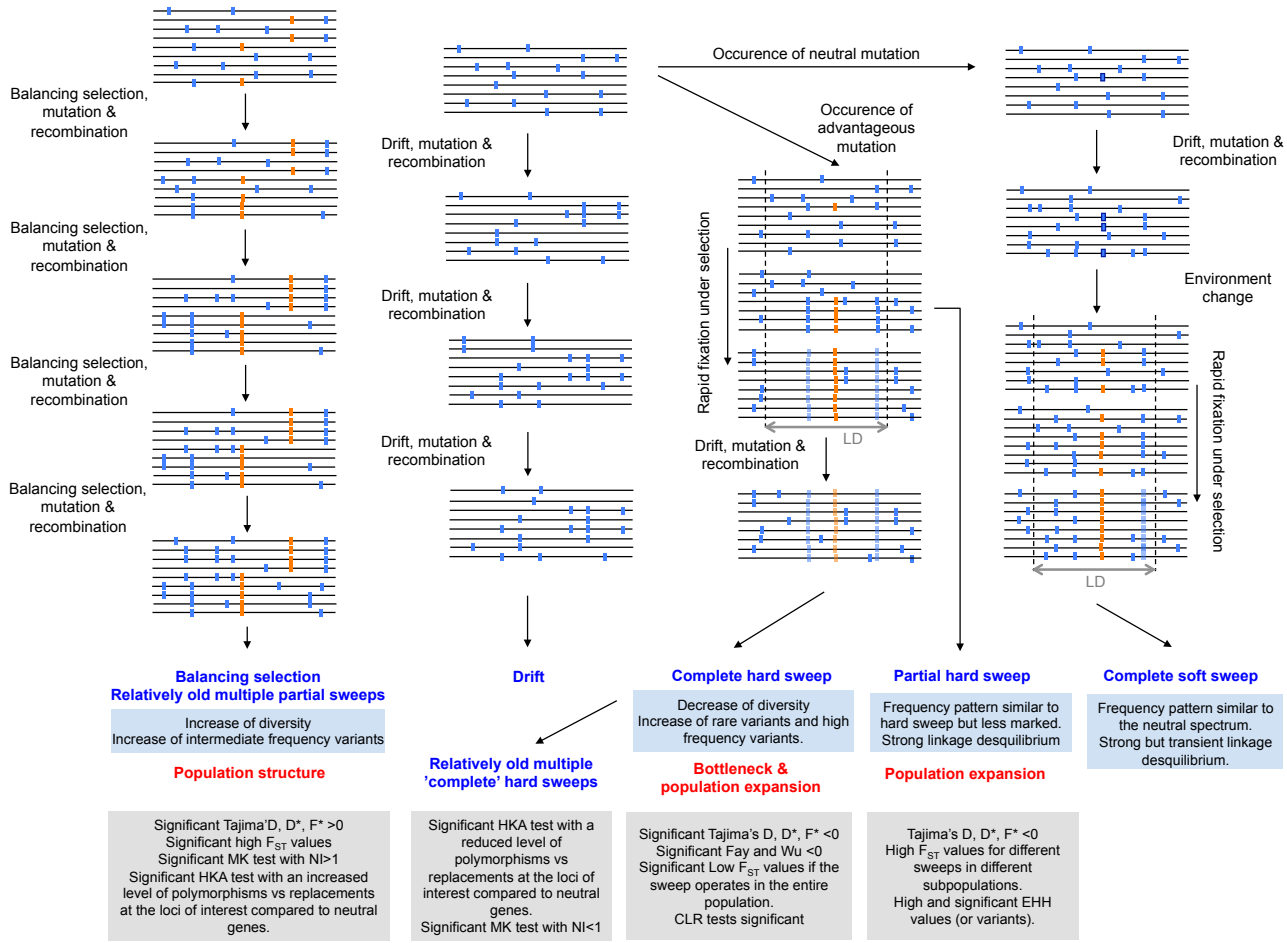


Fig. 20. Evolution of the patterns of polymorphisms expected at a loci under different modes of selection. Horizontal bars represent the different alleles of a chromosomic region segregating within a population. Neutral SNPs are indicated by small blue rectangle whereas advantageous SNP are in orange. Deleterious SNPs are not represented but could segregate more or less transiently depending on there effect on fitness and their genetic environment. The expected pattern of polymorphism and the expected values of relevant neutrality tests under each mode of selection are indicated in blue and grey boxes respectively. Possible confounding factors are indicated in red.

just experienced fixation (**complete sweep**) that will even be different from the one of old fixed variants (figure 20). Similarly, sweep can affect the whole species (complete sweep) or be restricted to a region or environment as a result of local adaption (partial sweep) and this implies difficulties when it comes to statistically recognizing those (figure 21).

Opposed to directional selection that favours one 'phenotype' and the underlying molecular variant, **balancing selection** actively maintain two (or several) alleles for a relatively long period of time in a population. This type of selection results from different processes including overdominance[12] (unlikely to be major in *A. thaliana*), frequency-dependent balancing selec-

12. Overdominance is observed when the heterozygote individual has higher fitness compared to either homozygote ($WAA < WAa > Waa$).

tion [13] and fluctuating selection coefficient in time or space that can cause, at the species or local scale, a rapid increase in frequency followed by either rapid loss or change under drift. The last point highlights the importance of the time and space scales at which we look phenomenon, as two old complete sweeps in two independent populations is a form of balancing selection (figure 21). Overall, balancing selection is often associated with the maintenance of old alleles that had more time to accumulate mutations and so results in an increase of genetic diversity and high frequency variants around the selected site (figure 20). The term of **diversifying selection** can also be used to refer to the maintenance of various adaptive alleles.

It is worth noting that all advantageous (deleterious) alleles are not fixed (removed) by selection. Indeed, the fate of a genetic variant also depends on its genetic environment and demographic context that can strongly influence its maintenance. Hitchhiking (or Hill Robertson effects) can be observed when neutral and slightly advantageous or deleterious alleles change in frequency as the result of their linkage with a variant under directional selection. Depending on the linked loci, slightly deleterious variants can be fixed with strongly positive linked variants (selective sweeps) or slightly advantageous loci counter-selected with strongly deleterious linked variants (background selection). In a population, the fixation of a positive variant can lead to the disparition of others slighty less advantageous variants (Hill Robertson interference). Overall, Hill Robertson effects reduce the efficacy of selection when recombination is low.

Finally, demographic factors can also leave signatures on the pattern of polymorphism. Overall, population expansion, bottleneck or structuration result in patterns similar to positive selection, hard sweep and balance selection respectively (figure 20). However demographic factors are likely to affect all the loci in the genomes whereas selection only affects punctual loci. This property is often used to test the significance of neutrality tests. However local variations in the mutation rate can also increase or decrease genetic variation at specific region but will have the same effect on both within and between species diversity. Besides, it won't modify alleles frequency spectrum.

3.2.2 Methods to detect evolutionary significant variants

As presented before, selection events, depending on their nature and age, result in different patterns of polymorphisms that are not expected under drift. There are roughly two kinds of tests that can be performed to detect molecular signal of directional positive or balancing selection. On the one hand, tests based on within-species diversity can help identifying ongoing or very recent selection events. On the other hand, tests based on between species divergence can detect repeated events of selection that took place over evolutionary time scale.

13. Frequency dependent balancing selection is observed when rare alleles have higher fitness than common alleles if they remain rare. As a result, rare alleles will be maintained within a population but as they increase in frequency, they will be less and less advantageous

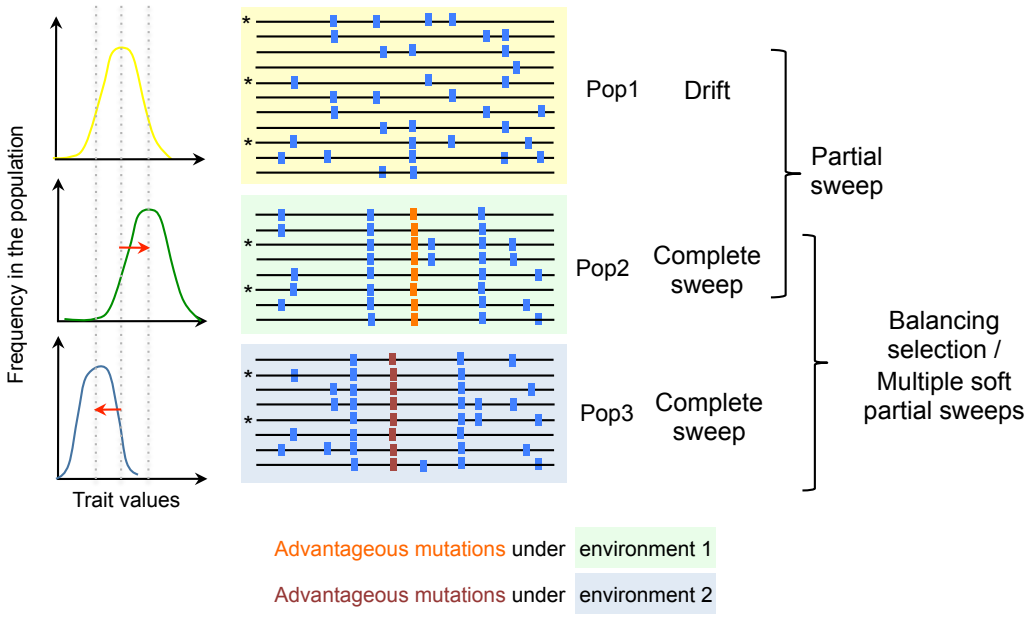


Fig. 21. Influence of the geographical scale on the observed pattern of polymorphism. Horizontal bars represent the different alleles of a chromosomic region segregating in 3 populations. Neutral SNPs are indicated by small blue rectangles whereas advantageous SNP in the green and blue environment are in orange and brown respectively. The phenotypic distribution of the trait of interest is represented for each population with red arrows corresponding to effect of adaptation in environments 1 and 2. At the local scale, the genetic region represented is either neutrally evolving (population 1) or have undergone independent sweeps due to the selection of two different variants in the population 2 and 3. Depending on the populations considered (braces) different pattern of polymorphisms can be observed at the species level and the selection acting at the loci can be seen as balancing selection (or multiple soft sweeps) or partial positive selection. If few individuals are sampled in the three populations (for example the ones indicated by $*$) the signal is likely to be confounded so that no selection event can be detected.

3.2.2.1 Tests based on within-species variation

First, we saw earlier that selection events affect differently the site frequency spectrum associated with a region i.e. the distribution across rare, intermediate and high frequency variants. Many of the classic neutrality tests try to capture this information by comparing different estimators of the scaled population mutation rate θ [14] . Under the equilibrium neutral [15] and infinite site [16] models, the number of segregating sites (S), the number of pairwise differences (π) and the number of singletons ($\eta*$) are a simple function of θ and should be consistent (table 3). The most famous test based on the frequency spectrum is the Tajima's D that compares the $\hat{\theta_\pi}$ with $\hat{\theta_S}$ (table 3) [244]. Whereas the number of segregating sites (S) only counts polymorphic positions independently of their frequencies and so is more sensitive to change in rare allele frequencies, the number of pairwise differences (π) accounts for allele frequencies and so is more sensitive to change in intermediate allele frequencies. A negative D-value indicates an excess of rare alleles which might appear under rapid directional positive selection, purifying selection or after a recent population expansion. A positive D-value indicates an increase of intermediate frequency variants which might appear under balancing selection or as the result of population subdivision [244] (figure 20). Two similar tests are the D* and F* tests developed by Fu and Li [246] although they consider the ancestral state of each mutation using an outgroup species and compare respectively $\hat{\theta_S}$ and $\hat{\theta_\pi}$ with $\hat{\theta_{\eta*}}$(table 3). Finally Fay and Wu [247] developed an estimator of θ that distinguishes ancestral and derived sites and accounts for high frequency derived alleles ($\hat{\theta_H}$). The test they propose compares $\hat{\theta_\pi}$ with $\hat{\theta_H}$ so that a negative value of the H test indicates an excess of high frequency derived variants compared to intermediate frequency variants excess that could transiently be observed after a sweep (table 3, figure 20).

One major problem with most neutrality tests is that they assume an equilibrium neutral model that is rarely plausible in natural populations. This is particularly true in *A. thaliana* which is mostly selfing, likely experienced recent expansion since the last glaciation and shows strong population structure (see 1.2.1.1). As a consequence, rejections based on comparisons with the theoretical distribution expected under the neutral model can be explained by both selection events or population demographic history. To circumvent this problem, one can compare the observed statistic to a distribution simulated using the coalescent theory [17] and taking into account major known demographic events (Ne, expansion, migration, bottleneck, structuration) [243, 248]. However this requires a good estimation of the demographic history of a species or

14. Under the neutral model $\theta = 4N_e\mu$ where N_e is the effective population size, μ is the rate of mutation per gene and per generation [245]

15. The equilibrium neutral model is valid for single panmictic diploid populations of constant and long term size and characterized by random mating. It stipulates that all mutations are neutral and appeared at a constant rate within the genome. This model does not allow migration, recombination or selection.

16. The infinite site model assumes that all mutations are neutral and create a new segregating site.

17. The coalescent theory is a popular probabilistic model that helps reconstructing individual genealogies based on DNA information.

Tab. 3. Different estimators of θ under the equilibrium neutral infinite site model based on four measures of variability as a function of the sample size n.

θ **estimator**	**Based on**	**Neutrality tests**
$\hat{\theta_w} = \frac{S}{a_n}$	S = number of segregating sites	Tajima's $D = \frac{\hat{\theta_\pi} - \hat{\theta_w}}{\sqrt{V(\hat{\theta_\pi} - \hat{\theta_S})}}$
$\hat{\theta_\pi} = \pi$	π = average number of pairwise differences	Fu and Li's $F* = \frac{\hat{\theta_\pi} - \hat{\theta_{\eta*}}}{\sqrt{V(\hat{\theta_\pi} - \hat{\theta_{\eta*}})}}$
$\hat{\theta_{\eta*}} = \frac{n-1}{n}\eta*$	$\eta*$ = number of singletons (using folded data)	Fu and Li's $D* = \frac{\hat{\theta_w} - \hat{\theta_{\eta*}}}{\sqrt{V(\hat{\theta_w} - \hat{\theta_{\eta*}})}}$
$\hat{\theta_H} = 2\sum_{i=1}^{n-1} \frac{i^2}{n(n-1)} n_i$	n_i = number of derived variants found i times in the sample	Fay and Wu's $H = \frac{\hat{\theta_\pi} - \hat{\theta_H}}{\sqrt{V(\hat{\theta_\pi} - \hat{\theta_H})}}$

See [242, 244, 246, 247] for details about the tests (variance (V)) – $a_n = \sum_{i=1}^{n-1} \frac{1}{i}$

population from genomic data. Another possibility is to use an outlier approach to compare the statistics observed at one gene to an empirical distribution obtained from genome-wide loci that have experienced past and ongoing demographic events. For example, in *A. thaliana*, Tajima's D statistics' distribution observed in 876 short DNA fragments of 96 accessions is skewed toward negative values (mean Tajima's D = -0.8 rather than the 0 expected under the neutral model) probably as a result of selection but also population expansion [14]. The problem of this approach is the estimation of the false positive rate from empirically observed outliers, a rather subjective choice. Besides, gene surfing, that can be observed when random alleles and mutations in the wave front of an expanding population surf to high frequencies, if it occurred, will likely result in outliers independent on the effect on fitness [249]. Theoretical and empirical distribution statistics can be analysed for most of the neutrality tests presented in this section.

Then, several methods have been developed to search for either, the localised depression of genetic diversity associated with sweeps or the localised increased variation associated with balancing selection (figure 20). Those methods go from unformal tests like the simple plot of genetic variation [168] as a function of map position, to more sophisticated likelihood-based tests that take into account the spatial pattern of variation to detect and localize sweep and infer the strength of selection such as the widely used CLR test (for composite likelihood ratio test) and its variants (Reviewed in [241, 242]). In *A. thaliana*, few strong and recent selective sweeps have been observed at the species level [35, 19] although several loci have been associated with high CLR values [250]. Overall, CLR tests are quite sensitive to demography;

more precisely to strong bottlenecks that can also lead to a strong decrease in genome-wide genetic variation. Besides, variations in mutation rate that can produce peak or valley of neutral variation and/or recombination rates are rarely taken into account but could strongly affect local frequency spectrum.

Whereas site-frequency spectrum tests ignore the associations among segregating sites (i.e. LD) some tests have been especially designed to analyse haplotypic structures. Those tests are based on the idea that frequent alleles are expected to be older under the neutral theory and so likely experienced more recombinations with tightly linked loci (reviewed in [242]). However, when a recent allele rapidly increases in frequency, recombination does not have time to reshuffle between tightly linked loci and so the region around the selected SNP exhibits a strong LD characterized by extended haplotypes (figure 20). Tests based on the detection of long haplotypes are rather robust to demography issues and are particularly powerful to detect partial sweeps. However they have little power to detect fixed hard sweeps as the region becomes highly homogeneous (there are too few individuals in each class of haplotypes to produce a meaningfull statistic). At the base of the most popular tests is the EHH value (extended haplotype homozygosity) that corresponds to the length of the region around a SNP of interest for which the probability that all the markers within the region are identical is 5% or greater. In *A. thaliana*, one of the variant of EHH that partially accounts for population structure has been widely used: PHS (Pairwise Haplotype Sharing) is an estimate of the shared length around any allele at a given position obtained from pairwise comparisons between individual haplotypes [251]. The association of this statistics with GWA or climatic information allows the detection of several partial sweeps likely evolutionarily important in *A. thaliana* [18, 250]. However, using another variant of EHH and a different experimental display, Fournier and colleagues did not detect any evidence that loci associated with high fitness were involved in partial sweeps [211].

Then, because population subdivision can result from adaptation to local environments [239], tests looking for high levels of population differentiation have been developed. Some of those methods used the famous FST value which is a measure of population structure and basically corresponds to the fraction of total genetic variance observed in a population that is due to genetic differences between population subgroups[18]. Excessively high values of FST at a marker locus indicate more divergence between subgroups than expected under drift, which might result from several positive directional selection events in different subgroups. Contrarily, low FST values indicate less divergence than expected under drift, which could be observed if balancing selection occured in several subgroups or in case of population-wide selective sweeps [241] (figure 20). In *A. thaliana*, analysis of FST fluctuations along the chromosomes allows the identification of a region responsible for the differentiation between European and Asi-

18. $FST = (\pi_{TOT} - \pi_{SUBPOP})/\pi_{TOT}$, but can also be measured as the proportion of reduction in heterozygosity for subpopulation compared with the total population $FST = (H_{TOT} - H_{SUBPOP})/H_{TOT}$.

atic/Russian populations [19].

Finally, some methods also account for the change in allele frequencies that can be observed under selection. Using temporal (series of time points) or spatial (a set of population) data, it is possible to compare observed changes in allele frequency with the one expected under genetic drift. Overall these methods are quite sensitive to population size and have low power to detect weakly-selected alleles [242].

Overall, the power of neutrality tests based on within-species genetic variation to detect selection events varies depending on the type of selection, the demographic history of the population (including changes in population size) and the local mutation and recombination rates. Because the tests presented above give different information about the pattern of polymorphisms and are differently affected by demographic events, combining different tests (preferentially uncorrelated ones) is important to get a better support for selection and reduce false-positives.

3.2.2.2 Tests based on between-species variation

Tests based on within-species variation can detect ongoing or very recent selective events but not sites that have experienced positive selection for a relatively high amount of time. To this end, two tests have been developed and assess whether the level of polymorphism within a population (or species) is consistent with the level of divergence between population (or species) across at least two categories of sites or genes, one neutral and one potential candidate for selection.

The first test was developed by Hudson, Kreitman and Aguadé (HKA) and compares the polymorphism/divergence ratio (P/D) over several loci (including likely neutral genes) [252]. Under the neutral model, the P/D is expected to be equal to $2N_e/t$ [19] where both Ne (the effective population size) and t (the divergence time) are the same for all genes. The P/D ratio among different genomic loci is expected to slightly vary due to variation in phylogenies or by chance and this variation can be estimated using coalescent simulations. Then using a χ^2 -like statistic the HKA test compare the P/D variation observed among loci with the one expected by chance. If the null hypothesis is rejected (i.e. if the observed P/D ratio varies too much among genes) an excess of polymorphisms within a species might be explained by balancing selection and a reduced level by positive directional selection. The HKA test have been widely used but variation in the recombination rates and demographic factors could also be responsible for variation in P/D ratio [241].

19. Under the neutral mode, $d_i = 2t\mu_i$ and $\theta_i = 4N_e\mu_i$ so that $\theta_i/D_i = 4N_e\mu_i/2t\mu_i = 2N_e/t$ where d_i is the divergence at gene i, θ_i the within population mutation rate or under the infinite site model the expected number of pairwise difference at gene i, t the divergence time, μ_i the mutation rate over the entire gene i and N_e the effective population size.

The second test, developed by McDonald and Kreitman (MK test) [253] compares the amount of divergence and polymorphism among synonymous polymorphisms and replacements. Under the neutral model, the ratio of non-synonymous to synonymous polymorphisms (P_{ns}/P_s) should be equal to the ratio of non-synonymous to synonymous fixed substitutions (D_{ns}/D_s). The number of polymorphisms/substitutions vs the type of changes (synonymous/replacements) can be associated in a simple contingency table and compared using a Fisher's exact test for the goodness-of-fit. The MK test is significant if P_{ns}/D_{ns} is significantly different from P_s/D_s, which can occur if there is either an excess of replacement polymorphisms (P_{ns}) (neutral index <1) or an excess of non-synonymous substitutions (D_{ns}) (neutral index >1) [20]. The former pattern is expected when positive selection has operated recurrently via the fixation of non-synonymous polymorphisms at a locus whereas the latter can be expected if balancing selection or several partial sweeps have been operating to maintain non-synonymous polymorphisms or if slightly deleterious replacement mutations have accumulated at low frequency. To partially solve the bias that can be observed due to slightly deleterious replacement mutations that often segregate at low frequency, it has been proposed to remove the low frequency variants from the analysis. However the threshold under which the polymorphisms have to be omitted is not clear [255, 256, 257]. Besides, removing the low frequency variants increases the number of false positives that can be observed due to slight increase in current population size compared to past population size [258]. Thus even though the MK test is considered as rather robust to demographic concerns, special care must be taken when interpreting a significant test. Overall every tests can give false-positives as a result of either population demography including gene surfing, variation in population size, bottlenecks and structuration or variation in local recombination or mutation rates. This is why population genetics analyses are not an absolute proof of selection but should be combined with functional and ecological analyses to clearly understand if a variant could be (or have been) adaptive in the wild.

3.3 Conclusion: Major evolutionary traits in A. thaliana?

The importance of flowering time in the adaptation of *A. thaliana* to natural environments makes no doubt. GWA and linkage mapping analyses identified many QTLs for flowering-related traits and several of them show correlations with climatic factors [73, 173, 224, 225, 226]. Genomic scans for selection identify a strong enrichment of SNPs associated with flowering related traits (GWA mapping) in the extreme tail of PHS and FST statistics distribution [18, 250] suggesting that several flowering QTLs are experiencing partial or ongoing sweeps as a result of differential selective pressures that could increase population differentiation. Consistent with this view, SNPs associated with high PHS statistic often show a narrow geographic distribution,

20. The neutral index proposed by Rand and Kann [254] is defined as followed: $NI = (P_{ns}/D_{ns})/(P_s/D_s)$.

a tendency that has also been observed for SNPs associated with several climatic variables such as day length [250] and for SNP associated with high fitness in common gardens [211]. Germination timing (and, hence, dormancy) is also an important phenological trait associated with fitness [212]. Overall, the climatic variables (i.e precipitation and temperature) that limit *A. thaliana* worldwide distribution range [8] are likely doing so by constraining fitness-associated phenological traits (flowering and germination timing) [73].

This conclusion does not mean that phenological traits are the main factors driving adaptation. Interestingly, Fournier-Level and colleagues [211] found several biotic- and abiotic-stress associated genes as strong candidates for local adaptation from associating climate differentiation, potential selective sweeps and high fitness in at least one common garden. Overall, the ecological relevance of genes involved in response to biotic and abiotic constraints is more difficult to assess because of the geographic heterogeneity of the underlying selective agents. Nevertheless, a recent study on *A. thaliana* local adaptation to herbivores using genotype-environment correlation and multigenerational selection experiments revealed that the variation in abundance of two specialist aphids could partially explain the geographical patterns of GS-ELONG chemotypes associated with the biosynthesis of different glucosinolates compounds [259]. This analysis does not only reveal the importance of herbivore communities in shaping *A. thaliana* local adaptation but also underlies the fitness cost associated with some defense mechanisms (here, the production of alkenyl glucosinolates) when the selective agent (here, leaf-chewing herbivores) is absent –a cost that has also been observed for other defense-related genes– [147, 260]. This cost/benefit balance is likely to play a major role in the maintenance of genetic variation observed at several R-genes [261, 262, 263]. Evolutionary analyses gave interesting insights into the modes of adaptation associated with defense-related traits. Overall, the evolutionary interaction between resistance genes in *A. thaliana* and pathogenic effectors from microbial pathogens is explained by both balancing selection and positive directional selection although little evidence of strong selective sweeps expected under the arm-races model[21] has been observed [18, 147, 264, 265]. Besides, it is not clear whether balancing selection results from frequency-dependent selection or from the adaptation to local pathogens by differential directional selection events in populations [240]. Finally, SNP associated to ionomic traits are significantly enriched in the extreme tail of the CLR scores distribution. As a result, complete hard sweeps may be one mode of adaptation for several ionomic traits (boron concentration for example) [18]. However, because of the heterogeneity of edaphic parameters and because ionomic traits are controlled by many loci of relatively small effects, it is likely that balancing selection, partial and soft sweeps also contribute to the evolution of ionomic traits [145]. Candidate genes responsible for the natural variation of *A. thaliana* response to abiotic constraints are still lacking compared to phenological traits, which make the comparison of the evolution

21. The 'arm-race' model stipulates that selective sweeps favour alternate host and pathogens changes.

mechanisms acting on those traits difficult. Overall, the importance of each mode of selection regarding the adaptation of *A. thaliana* to the environment and the evolution of *A. thaliana* genome diversity remains elusive.

4. SCOPE OF THE THESIS

My doctoral work focuses on analysing natural variation for shoot growth and response to the environment in *A. thaliana*. As presented before, natural variation analyses aim at understanding how molecular genetic or epigenetic diversity controls phenotypic variation at different scales and times of plant development and under different environmental conditions, and how selection or demographic processes influence the frequency of those molecular variants in populations for them to get adapted to their local environment. As such, the analysis of *A. thaliana* natural variation can be addressed using a variety of approaches, from genetics and molecular methods to ecology and evolutionnary questions. During my PhD, I got the chance to tackle several of those aspects through my contributions to three independent projects which have in common to exploit *A. thaliana* natural variation.

My first contribution consisted in the validation of the non-neutral evolution of a molybdenum transporter encoding gene, *MOT1*. Analysis of shoot growth natural variation under acidic conditions by Dr. Poormohammad led to the identification of a new hypofunctional variant at MOT1. Although associated with low Mo content and growth defects under Mo deficiency, this variant is frequent in West-Asia populations in which allelic variation is strictly correlated with the concentration of available Mo in native soils. I demonstrated that Mo could be toxic at high concentration suggesting a possible adaptive role of MOT1 hypofunctional variant for Mo toxicity tolerance. Supporting this hypothesis, using a population genetics approach including a combination of different neutrality tests, I showed at different geographical scales that *MOT1* pattern of polymorphisms in *A. thaliana* was not consistent with neutral evolution and more likely supported that diversifying selection was acting at this locus.

My second contribution consisted in the functional characterisation of two SD1 receptor-like kinases responsible for shoot growth variation under mannitol stress in *A. thaliana*. When I arrived at Dr. Loudet's lab, a shoot growth QTL isolated in a recombinant inbred line (RIL) population derived from Col-0 and Cvi-0 accessions had been fine mapped to a 10kb region encompassing 3 genes. First, I showed that the best quantitative trait gene (QTG) candidate, annotated on TAIR10 as encoding a double receptor-like kinase (RLK), actually correspond to two different transcriptional units (renamed *At1g11300* and *At1g11305*) that appeared by tandem duplication. Using different genetic tools (mutant analysis, transgenics and specific association genetics), I showed that those RLK were necessary to induce a specific response

on mannitol-supplemented media that was not due to the osmotic properties of mannitol but rather to the molecule itself. This result raised the question of the function of such receptors in *A. thaliana*, which does not synthesize mannitol. Our hypothesis is that the activation of those RLKs by the mannitol produced and released in planta by some pathogens such as fungi, could participate to plant defence. This work also implies that some confusion between biotic and abiotic responses could have arisen from the utilization of mannitol as an osmoticum.

Finally, in parallel to these two main projects and in collaboration with Dr. Michel Vincentz's group from the Center for Molecular Biology and Genetic Engineering (Brazil), I was involved in the characterisation of natural epialleles at *QQS*, an *A. thaliana* de novo originated gene known to regulate starch metabolism. While analyses of *A. thaliana* epigenome suggested that *QQS* was under epigenetic control, Dr. Amanda Silveira unexpectedly identified a range of spontaneous epiallelic stable variants in a Col-0 laboratory seed stock whose *QQS* basal expression level was strictly correlated with the methylation of its promoter and 5'UTR regions. I contributed to the confimation that the methylation status of *QQS* cis-acting regulatory elements was the main element explaining *QQS* transcript accumulation differences in different genetic backgrounds. Beside, I showed that *QQS* epialleles were not the result of artificial lab conditions and were really segregating in natural populations facing selective pressures. Finally, within one of the west-Asian population, I identified two genetically distinct (but closely related) subpopulations that differed in their *QQS*-methylation patterns. This latter discovery raised the question of the origin and dynamic of the different *QQS* epivariants in wild populations.

In the next three parts of this manuscript – each part corresponding to one project – I will first briefly introduce the project context, the main results and clarify my contributions. Then the obtained results are detailed in a scientific paper that has either been accepted or is about to be submitted to an international peer-reviewed journal. Finally, I use additional results, that couldn't be published because negative, too long or because they would require further investigations, to discuss in more details some of the results presented in the scientific paper. I conclude on the importance and complexity of the effects of environmental parameters on phenotypic traits.

Part II

ALLELIC HETEROGENEITY AND TRADE-OFF SHAPE NATURAL VARIATION FOR RESPONSE TO SOIL MICRONUTRIENT

5. PROJECT BACKGROUND AND PERSONAL CONTRIBUTION

The '*Tourbe*' project started in 1999. Back then, Olivier as a PhD student was looking for a non-enriched compost to study *A. thaliana* natural variation for growth-related traits in interaction with nitrogen availability. He started working with peatmoss, a poor acidic substrate, and observed that, compared to Bay-0 (from Bayreuth, Germany), the Shahdara accession (from the Shakhdara valley, Tajikistan) grew very poorly on this substrate. After his postdoc, back in Versailles, he decided to map the loci responsible for this growth difference. The mendelian segregation of the character among Bay-0 x Sha RILs indicated that one major locus was controlling the shoot growth phenotype. Olivier's technical assistant, Claire Le Metté and later on his postdoc, Kian Poormohammad, mapped this locus on chromosome 2 and reduced the candidate interval to 80kb. Among the 18 genes present in this interval, *MOT1*, identified by two independent groups as responsible for quantitative variation in molybdenum (Mo) content in the shoot and root of different Arabidopsis accessions, was a good candidate.

Mo is an essential micronutrient for prokaryotes and eukaryotes. It is the cofactor of several enzymes that catalyze important oxido-reduction reactions. In plants, identified molybdoenzymes are the nitrate reductase (NR), the aldehyde-oxidase, the sulfite reductase and the xanthine oxidase. These enzymes are respectively involved in various and important developmental processes such as nitrate assimilation [266], hormones (abscisic acid (ABA) and auxin) and glucosinolate synthesis [267, 268, 269, 270], sulfite detoxification [271] and purine degradation [272]. Plants affected in the biosynthesis of pterin-based Mo cofactor (moco) are lethal when grown in soils, like NR mutants. Besides, plants grown under Mo deficiency are strongly affected in their growth with lesions, low chlorophyll levels and low ABA contents [273]. In soils, Mo exists under various forms including mineral forms ($MoSe$, $PbMoO_4$, $Fe_2(MoO_4)$) and complexes with other elements but only soluble molybdate (MoO_4^{2-}) can be taken up by plants [274, 266]. Under low pH (<5) Mo is adsorbed by soil colloids, so that less molybdate is available for plants and they may then present the Mo-deficiency phenotypes described above.

In archae and eubacteria, molybdate is transported through a well described ABC-type transporter or unspecific anion carrier proteins [275]. In plants, two specific molybdate transporters have been described. *MOLYBDATE TRANSPORTERS MOT1* and *MOT2* genes (previously described as *SULTR5.1* and *SULTR5.2* [276]), are close paralogs belonging to the fifth

group of the SULTR family that comprise putative sulphate transporter. Mutation in *MOT1* (*mot1.1*) results in low Mo levels in roots and shoots resulting in reduced growth and lesions under low Mo availability [144, 143]. However the low shoot Mo level observed in *mot1* mutant is likely driven by the root [143]. MOT1 specifically transports Mo and not sulphate [144, 143] and this gene is expressed in roots and shoots. Its cellular localisation is unclear. Tomatsu et al. observed MOT1 in the endomembrane system whereas Baxter et al. observed the transporter at the mitochondrial membrane. The N-terminus GFP fusion created by Tomatsu et al. could block the mitochondrial signal peptide of MOT1 likely resulting in mislocalization [144, 143]. Overall the role of MOT1 remains obscure. It could regulate whole plant Mo accumulation at the level of the mitochondria in the roots. MOT2 molybdate transport activity has not been demonstrated directly. Neverthless, MOT2 is clearly involved in the remobilization of Mo from senescing leaves into developing seeds as reflected by the *mot2* mutant phenotype that presents increase and decrease levels of Mo in senescing leaves and seeds respectively. *MOT2* encodes a vacuolar transporter and its expression increase in leaves during senescence [277]. Finally, *MOT1* has been identified as a major loci controlling Mo accumulation in *A. thaliana* in several independent linkage mapping [144, 143, 166, 92] and one GWA mapping analyses [74]. A 53-bp deletion in the *MOT1* promoter was observed in several accessions including Ler and Van-0, and lead to a strong reduction of *MOT1* transcript accumulation in roots and shoots. This deletion is likely responsible for the low the Mo contents observed in those accessions.

Going back to '*Tourbe*' project, as low pH conditions affect Mo availability, Kian and Olivier hypothesized that the Shahdara phenotype observed on peatmoss could result from low Mo availability. Consistent with this hypothesis, Kian observed that the growth of *mot1.1* was affected on peatmoss compared to Col-0. Besides, quantitative complementation of the *mot1.1* defective allele with both Col and Bay alleles, but not with Sha allele, confirmed MOT1 as the causal gene and the defectiveness of Sha allele. Finally, adding Na_2MoO_4 to the watering solution complemented the phenotype of Shahdara and *mot1.1* on peatmoss, without affecting the pH of the soil. When Kian sequenced *MOT1[Bay]* and *MOT1[Sha]* alleles, he identified one non-synonymous polymorphism D104Y, that was very likely the polymorphism causing Sha-phenotype on peatmoss. To link the growth defect on peatmoss and the presence of either the D104Y mutation or the 53-bp deletion identified by Tomatsu and Baxter, Kian screened 299 accessions for the presence of those polymorphisms. He found 32 accessions with the D104Y mutation (Sha-like accessions) and 21 accessions with the 53-bp deletion (Ler-like accessions). All Sha-like and Ler-like accessions tested presented a growth defect on peatmoss (figure 22). Thus, he provides here a clear example of allelic heterogeneity with at least two different alleles of the same gene resulting in a similar phenotype.

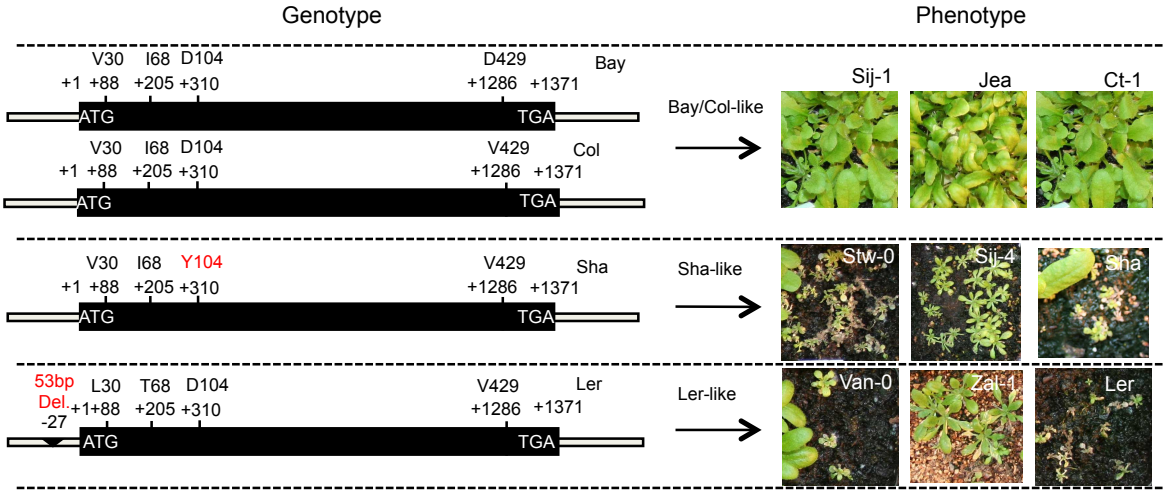


Fig. 22. Allelic heterogeneity at *MOT1* in *A. thaliana*. Two independent alleles (in red) are responsible for shoot growth defects when accessions are grown on peatmoss (here 35 DAS).

When I arrived in the lab, the evolutive relevance of *MOT1* allelic heterogeneity was unknown. During my master and my PhD, with the help of Matthieu Simon and then Thierry Robert, I performed population genetics analyses on the *MOT1* gene to detect possible traces of selection (Publication – TableS1, S2, S3 – table 4, table 5). Our results combined to the observation that *MOT1[Sha]* allele is preferentially associated with Mo-rich soils suggested that this allele could have been selected in 'West Asia' to face high Mo-toxicity. Then I participated to the confirmation of that hypothesis by testing the effect of high Mo on the growth and fitness pattern of various *A. thaliana* accessions and mot1.1 mutant (Publication –Fig. S6 – figure 28). Finally, I also performed some experiments that were necessary for the paper but not done by Kian before he left like qPCR experiments that confirmed that *MOT1[Sha]* was expressed at a level similar to *MOT1[Bay]* (Publication –Fig. S3–). It is still a rare finding to be able to relate functional genetic variants to fitness or trade-off effects and even more to associate this variation to the environment. This work indicates that environmental parameters of importance, such as soil properties, may be heterogeneously distributed and therefore require local description and study of local adaptation, which is greatly facilitated by the identification of the causative locus. Yet, our results also highlight the difficulty in formally testing adaptive hypotheses in genetic backgrounds and environmental conditions that are not exactly what exists in the wild.

6. PUBLICATION IN PLOS GENETICS

ALLELIC HETEROGENEITY AND TRADE-OFF SHAPE NATURAL VARIATION FOR RESPONSE TO SOIL MICRONUTRIENT.

The '*Tourbe* ' project has been published in PLoS Genetics in July 2012. My contributions to this paper are highlighted in the figures 4, S3, S6, S7 and in the tables S1, S2 and S3.

OPEN ACCESS Freely available online

PLoS GENETICS

Allelic Heterogeneity and Trade-Off Shape Natural Variation for Response to Soil Micronutrient

Seifollah Poormohammad Kiani[1], Charlotte Trontin[1], Matthew Andreatta[2], Matthieu Simon[1], Thierry Robert[3], David E. Salt[2], Olivier Loudet[1]*

1 INRA, UMR1318, Institut Jean-Pierre Bourgin, Versailles, France, **2** Department of Horticulture and Landscape Architecture, Purdue University, West Lafayette, Indiana United States of America, **3** Laboratoire d'Ecologie, Systématique, et Evolution, Université Paris-Sud XI, Orsay, France

Abstract

As sessile organisms, plants have to cope with diverse environmental constraints that may vary through time and space, eventually leading to changes in the phenotype of populations through fixation of adaptive genetic variation. To fully comprehend the mechanisms of evolution and make sense of the extensive genotypic diversity currently revealed by new sequencing technologies, we are challenged with identifying the molecular basis of such adaptive variation. Here, we have identified a new variant of a molybdenum (Mo) transporter, MOT1, which is causal for fitness changes under artificial conditions of both Mo-deficiency and Mo-toxicity and in which allelic variation among West-Asian populations is strictly correlated with the concentration of available Mo in native soils. In addition, this association is accompanied at different scales with patterns of polymorphisms that are not consistent with neutral evolution and show signs of diversifying selection. Resolving such a case of allelic heterogeneity helps explain species-wide phenotypic variation for Mo homeostasis and potentially reveals trade-off effects, a finding still rarely linked to fitness.

Citation: Poormohammad Kiani S, Trontin C, Andreatta M, Simon M, Robert T, et al. (2012) Allelic Heterogeneity and Trade-Off Shape Natural Variation for Response to Soil Micronutrient. PLoS Genet 8(7): e1002814. doi:10.1371/journal.pgen.1002814

Editor: Rodney Mauricio, University of Georgia, United States of America

Received April 6, 2012; **Accepted** May 21, 2012; **Published** July 12, 2012

Funding: The authors received no specific funding for this work.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: Olivier.Loudet@versailles.inra.fr

Introduction

Some of the most important constraints that plants have to adapt to are those related to soil properties [1,2]. These are also possibly some of the least well studied constraints, because they are spatially heterogeneous, thus not prone to typical geographic clines [3,4], and require analysis at the local/population scale [5]. In this context, quantitative genetics approaches hold great promise to reveal the genetic basis of adaptation by enabling the identification of the molecular origin of phenotypic differences between populations or even between species [6,7]. One of the benefits of identifying the causative polymorphism(s) and/or gene(s) explaining natural phenotypic variation is that it allows direct testing for correlations between environmental factors populations may be responding to and the occurrence of the target genetic polymorphism. This contrasts with working indirectly through populations phenotype, which may reflect contradictory patterns and trade-offs, if not genetic drift [5,8]. Moreover, this approach also enables direct testing for fitness advantages or potential cost of adaptation (i.e. antagonistic pleiotropy, an expected argument for local adaptation), that may be masked by linkage to deleterious mutations or genetic drift in reciprocal transplant experiments [9,10]. Molecularly identified examples of potentially-adaptive variation are still largely lacking and the debate is open as to the scale and rate of adaptive evolution [11,12].

In this work, we aimed at identifying the molecular bases of natural variation in accumulation of an essential micronutrient and understanding the ecological significance of this diversity. We describe both the fitness trade-offs of this variation and its potential adaptive advantage in the environment, revealing a system that is unlikely to have remained neutral.

Results/Discussion

Bay-0 and Shahdara, two strains (accessions) derived from wild populations of *Arabidopsis thaliana*, show contrasted growth behavior when grown on acidic peatmoss substrate (Figure 1). Using a segregating population derived from the cross of these two accessions, we determined the Shahdara growth defect to be segregating from a major-effect recessive locus on chromosome 2, as confirmed by a near-isogenic line derived from a residual heterozygous interval in one of the recombinant inbred lines (HIF084; Figure 1). Additional recombinant lines were phenotyped and genotyped to pinpoint the causative interval to 80 kb (Figure S1), covering 19 annotated genes. Among these, one gene appeared as a good functional candidate: *MOT1* was previously linked to the transport and homeostasis of the essential micronutrient molybdenum (Mo) in the plant [13,14], an element which availability is known to vary with soil pH [15]. Indeed, a T-DNA insertion mutant in the *MOT1* gene (*mot1.1*; Figure S1) shows a phenotype similar to Shahdara on peatmoss and –contrary to the Bay and Col alleles– the Sha allele is not able to restore wild-type growth when combined with the mutant allele in F1 hybrids (Figure 2). Moreover, we show that defective growth is complemented by either increasing soil pH with additional $CaCO_3$ mixed to the peatmoss (Figure S2) or increasing Mo availability (without altering soil pH) by adding Mo in the watering solution (Figure 3). Hence, genetic and chemical complementation shows that acidic

Author Summary

Plants are studied for their ability to adapt to their environment and especially to the physical constraints to which they are subjected. It is expected that they evolve in promoting genetic variants favorable under their native conditions, which could lead to negative consequences in other conditions. One approach to study the mechanisms and dynamics of these adaptations is to discover genetic variants that control potentially adaptive traits, and to study directly these variants in wild populations to try to reveal their evolutionary trajectory. We have identified a new polymorphism in a gene coding for a transporter of molybdenum (an essential micronutrient for the plant) in Arabidopsis; we show that this variant has strong phenotypic consequences at the level of plant growth and reproductive value in specific conditions, and that it explains a lot of the species diversity for these traits. Especially, the variant is associated with a clear negative effect under molybdenum-deficient conditions (caused by soil acidity) and with a subtle positive effect under molybdenum-plethoric conditions. Interestingly, the landscape distribution of the variant is not random among Asian populations and correlates well with the availability of molybdenum in the soil at the precise location where the plants are growing in the wild.

soil pH is responsible for reducing Mo-bioavailability and that, combined with a defective allele at *MOT1*, this results in the typical Mo-deficiency syndromes of reduced leaf Mo contents, strongly altered growth and development, necrosis [16]. These observations of a significant phenotypic consequence of variation at *MOT1* provide a model for the potential adaptive significance of this variation that goes beyond the simple variation in Mo content revealed previously [13,14].

Although we find that Landsberg *erecta* (L*er*) has a similar behaviour than Shahdara in our conditions (Figure 3), this defective allele ($MOT1^{Ler}$) used initially to reveal the gene's activity [13,14] is functionally different from the $MOT1^{Sha}$ defective allele. $MOT1^{Sha}$ doesn't bear the promoter 53 bp-deletion as in L*er* (Figure S1) and in fact is not showing $MOT1^{Ler}$-like transcriptional down-regulation compared to $MOT1^{Bay}$ or $MOT1^{Col}$ (Figure S3). Instead, $MOT1^{Sha}$ seems defined by a single amino-acid change in the protein relative to Bay-0 and Col-0 (Figure S1), strongly suggesting that $MOT1^{Sha}$ is hypofunctional. However, the MOT1 protein produced from the Sha allele is still able to increase Mo accumulation when heterologously expressed in yeast (Figure S4).

We then genotyped a random worldwide sample of ~300 accessions for the Sha-like amino-acid change and the L*er*-like 53-bp deletion and find that these alleles are both present at intermediate frequencies (15–20%) among the populations. Sequencing 102 of these accessions for the whole gene and promoter region revealed that the very conserved $MOT1^{Sha}$ haplotype is indeed clearly defined solely by the D104Y amino-acid change, while the $MOT1^{Ler}$ genotype is more complex and diverse (Table S1). All Sha-like and L*er*-like accessions that have been phenotyped show that both $MOT1^{Sha}$ and $MOT1^{Ler}$ haplotypes are perfectly associated with defective growth under acidic soil conditions (Table S1) and complementation crosses with five additional Sha-like accessions confirm allelism to *mot1.1* (Figure S5). Taking into account this allelic heterogeneity now explains most of the species variation toward low-Mo contents revealed in previous work [13] (http://www.ionomicshub.org/arabidopsis/). This form of complexity –in addition to genetic heterogeneity– is probably more frequent than previously thought in many organisms and is likely to help explain part of the missing heritability [17,18].

Regarding *MOT1* defective haplotypes, $MOT1^{Sha}$ is confined to 'West-Asia' (including Russia) with a high frequency among these populations (Figure S6) and displays a very low polymorphism level ($\pi Sha = 0{,}00016$) in comparison to other haplotype clusters, including the worldwide-distributed $MOT1^{Ler}$ allele ($\pi Ler = 0{,}0017$; Table S1). This may translate a recent and rapid expansion of the Sha allele through 'West-Asia', which could be due to neutral processes such as gene surfing associated with post-glaciation recolonization events from Central Asia [19]. This may also witness local positive selection events in favour of the Sha allele. Indeed, patterns of nucleotide polymorphisms at *MOT1* in the sample of 102 accessions strongly deviate from the expectation under the strict neutrality model, contrarily to two control loci, *PI* and *COI1* (Table S2). Negative values of Tajima's D reveal an excess of rare alleles at *MOT1*, suggesting the possible occurrence of at least one past selective sweep that has targeted this locus. Other well documented evolutionary processes such as population expansion after the last glaciation event [20] and population genetic structure [21] could also have contributed to the excess of rare alleles observed at the genome-wide level [22], as well as at the *MOT1* locus in *A. thaliana*. Nevertheless, the HKA and McDonald-Kreitman tests, which do not rely on the frequency spectrum, support the hypothesis of diversifying selection at the species level (Tables S2 and S3). The excess of within-species polymorphisms relatively to inter-specific divergence and the excess of non-synonymous polymorphisms observed at *MOT1* may result from the selection of different haplotypes at the worldwide scale. Interestingly, this trend is also clear when considering only accessions from 'West Asia', suggesting that the selection process could happen at different geographical scales.

Our own documented collection of wild populations from diverse regions in 'West-Asia' allowed us to investigate potential relationships between *MOT1* alleles and environmental parameters described precisely at the population site, especially soil properties. We saw no relationship with soil pH (indeed, none of the described populations were facing acidic soil conditions), but there was an obvious trend for populations with the defective $MOT1^{Sha}$ allele to grow on soils with high water-extractable Mo content (Figure 4). This may indicate that the defective Sha allele is a protective response to Mo accumulation in environments with excess Mo. Indeed, under such conditions in the laboratory, we observe a strong decrease in fitness (through the total number of seeds produced per plant) in all genotypes (Figure S7), indicating that plants have to find the right balance between Mo deficiency and Mo toxicity, a trade-off that could be resolved partly through variation in function of the Mo transporter MOT1. Moreover, we show that a defective *MOT1* allele (either *mot1.1* or $MOT1^{Sha}$) is accompanied by a slightly increased average seed mass (another component of fitness) specifically under Mo-toxic conditions (Figure S7), the outcome of which is difficult to estimate in nature [23]. It is however worth noting that previous studies in *A. thaliana* have shown positive effects of increased seed size for example on subsequent root and shoot growth [24] or seedling survival under limiting conditions [25].

In summary, we have identified a new functional variant at *MOT1* that contributes to explain most of the species' diversity in Mo homeostasis, and associated phenotypes that provide likely explanations for its non neutral evolution and its correlation to native soil. It is still a rare finding to be able to relate functional genetic variants to fitness or trade-off effects [26–28], and even more to associate this variation to the environment [3,7]. Our work indicates that environmental parameters of importance, such as soil properties,

Figure 1. Acidic peatmoss substrate induces severe growth defect linked to the Shahdara allele. When grown on peatmoss at a pH close to 5, Shahdara is subject to severe growth and developmental arrest, necrosis and death, contrary to Bay-0 which develops normally. In the cross between these two strains, this phenotype is entirely controlled by a single locus (Figure S1), as confirmed by near-isogenic line 'HIF084' segregating solely for a region of chromosome 2.
doi:10.1371/journal.pgen.1002814.g001

may be heterogeneously distributed and therefore require local description [5] and study of local adaptation [9,11], which is greatly facilitated by the identification of the causative locus.

Materials and Methods

All accessions used and the Bay-0×Shahdara RIL set [29] were obtained from Versailles Arabidopsis Stock Centre (http://dbsgap.versailles.inra.fr/vnat/). Heterogeneous Inbred Family 'HIF084' was derived from RIL084 (segregating for the region of interest) as previously described [30]. New collections of *A. thaliana* accessions were partly described previously [31] and are shown at http://www.inra.fr/vast/collections.htm. T-DNA insertion mutant *mot1.1* corresponds to line SALK_118311 as described [13,14]. Genetic complementation tests were performed on F1 plants issued from the cross of diverse accessions to *mot1.1* or its wild-type background.

Acidic soil assays were performed on 'Floratorf' peatmoss (Floragard, Germany) mixed with $CaCO_3$ (4 g per liter of dry peatmoss) to maintain a soil pH~5, watered with classical nutrient solution and grown under typical long-day conditions at 20°C. Chemical complementations were achieved in the same condition but, either with 8 g $CaCO_3$ per liter of peatmoss to reach a pH~6, or using a watering solution supplemented with 1 mM Na_2MoO_4. Mo toxicity was tested on regular fertilised soil mix (pH = 6) watered with nutrient solution supplemented with 7 mM Na_2MoO_4, or not (control).

MOT1 sequencing, qPCR analysis of expression (normalised against *GAPDH* and *PP2A*), functional characterisation in yeast were performed as previously described [13]. Extractable Mo was determined in soils by the method of Soltanpour and Schwab [32] using ICP-MS as the detector.

A total of 102 *A. thaliana* accessions (including 48 accessions known to maximize *A. thaliana* diversity [33] and 44 accessions

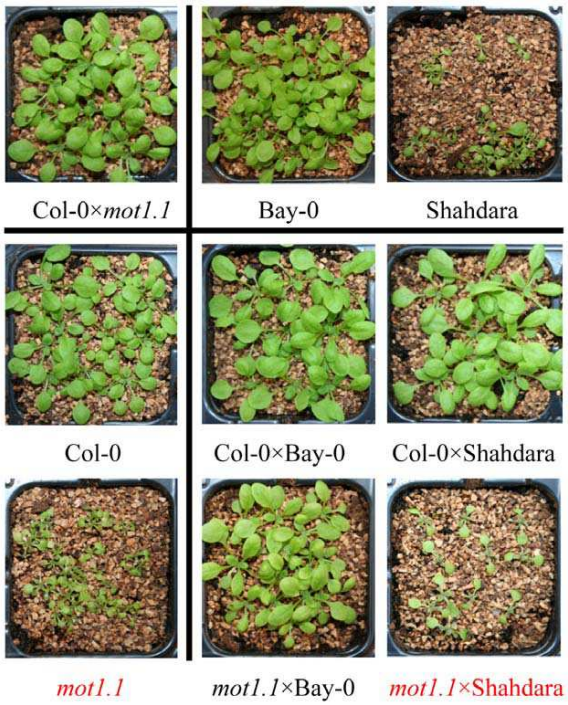


Figure 2. Mutant analysis and allelic complementation confirms *MOT1* as the likely causative gene. Peatmoss phenotype of diverse genotypes is shown, including Bay-0 and Shahdara parental lines, *mot1.1* mutant and its wild-type genetic background (Col-0), and F1 plants from complementation crosses between these genotypes. Unlike other alleles, the Shahdara allele at *MOT1* is not able to rescue the mutant phenotype.
doi:10.1371/journal.pgen.1002814.g002

from 'West-Asia') and 5 *A. halleri* accessions (I-14, I-16, F-1, PL-22 and TZC; obtained from H. Frérot at Univ. Lille [34]) were sequenced at *MOT1* (including 1 kb upstream and 0.3 kb downstream for *A. thaliana* accessions) and at two reference loci, *COI1* (At2g39940; 2,600 bp coding sequence) and *PI* (At5g20240; 2,150 bp coding sequence). Those genes were either used previously as reference or shown to have a neutral pattern of polymorphisms in *A. thaliana* [35,36]. Sequences were aligned using Codoncode Aligner v3.7.1. and subsequent alignments were improved visually.

Intraspecific analyses *i.e.* nucleotide diversity estimated by π [37] and θ_w [38], and Tajima's D statistics [39] were calculated using DNAsp v5.10.01 on the whole region sequenced. Ten thousands coalescent simulations under the strict Wright-Fisher neutral model assuming no recombination and conditioning on S were performed to estimate statistical significance of Tajima's D.

For interspecific analyses, the orthologous of *MOT1* and of the two control loci in *A. halleri* were used. The McDonald-Kreitman test [40] was performed by using DNAsp v5.10.01 in order to test for possible excess or deficiency in replacement substitutions at *MOT1*. Singletons were discarded for this analysis in order to reduce the contribution of slightly deleterious mutations (expected at very low frequencies and unlikely to become fixed). Neutral index was calculated as previously described [41]. Divergence between *A. thaliana* and *A. halleri*, defined as the average number of nucleotide differences between populations per gene, was calculated using DNAsp v5.10.01 and used to perform HKA tests with the multilocus HKA program available from J. Hey laboratory (http://genfaculty.rutgers.edu/hey/software).

Supporting Information

Figure S1 Fine-mapping the causative locus identifies *MOT1* as a candidate gene. A. The physical region of chromosome 2 found to be linked to the growth defect phenotype is shown (physical positions are given in Mb). B. Zooming in the candidate region highlights recombinants within the inbred lines that allow to fine-map the causative locus, thanks to additional markers (vertical dashed lines). Individual lines' genotype are depicted in horizontal coloured boxes (green = Bay allele; purple = Sha allele ; dashed = -heterozygous) and their phenotype on peatmoss are indicated (S = Sensitive; R = Resistant). C. This allows to narrow down the causative region to 80 kb, a region containing 19 predicted genes including the candidate At2g25680 (*MOLYBDATE TRANSPORTER 1*). D. *MOT1* has been sequenced in parental accessions and polymorphisms between Col-0, Bay-0 and Shahdara are represented along the single-exon gene model, including an amino-acid change specific to Shahdara (D104Y). The position of the insertion of a T-DNA in the *mot1.1* mutant (SALK_118311) is indicated.
(TIF)

Figure S2 Chemical complementation links growth defect with soil pH. Increasing soil pH from ~5 ('Control') to ~6 ('+$CaCO_3$') by doubling the amount of $CaCO_3$ mixed to the peatmoss substrate rescues normal vegetative growth of the Shahdara strain.
(TIF)

Figure 3. Chemical complementation links growth defect with Mo shortage. Accessions with potentially functional (Bay-0, Col-0) and defective *MOT1* alleles (*Ler*, Shahdara, *mot1.1*) were grown on peatmoss substrate watered with nutrient solution containing either traces of Mo ('Control') as in Figure 1, or 1 mM Na_2MoO_4 ('+Na_2MoO_4'). pH was checked to remain unchanged across treatments at ~5.
doi:10.1371/journal.pgen.1002814.g003

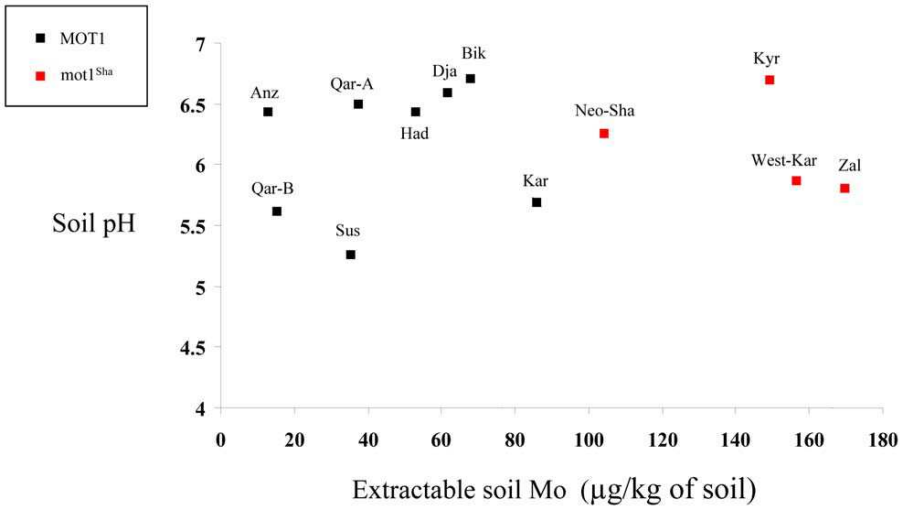


Figure 4. West-Asian populations highlight the correlation between *MOT1*^{Sha} allele and high native soil Mo content. Soil parameters (pH, extractable Mo) from the precise original collection site are represented for independent populations carrying (red dots) or not (black dots) the *MOT1*^{Sha} allele.
doi:10.1371/journal.pgen.1002814.g004

Figure S3 *MOT1* transcript accumulation does not explain *MOT1*^{Sha} defective allele. *MOT1* transcript accumulation relative to *GAPDH* and *PP2A* controls is shown from roots of diverse genotypes as in Figure 3. Contrary to L*er* and *mot1.1*, Shahdara accumulates normal levels of transcript. Standard errors are shown.
(TIF)

Figure S4 MOT1^{Sha} is able to transport Mo in yeast. The Sha allele of MOT1 was overexpressed in yeast heterologous system (Sha/p416) and shown to lead to Mo accumulation compared to the empty vector (p416, used as reference) or the yeast wild-type strain (BY4741), to an extent not significantly different from the Col MOT1 allele (Col/p416). Error bars represent interquantile range of medians.
(TIF)

Figure S5 Multiple accessions sharing the *MOT1*^{Sha} haplotype confirm the causative gene and polymorphism. Peatmoss phenotype of diverse genotypes is shown: 5 independent Sha-like accessions (Stw-0, Kly-2, Sij-4, Kondara and Zal-3) and F1 plants from complementation crosses between each of these accessions and either the *mot1.1* mutant or its wild-type genetic background (Col-0). As in Figure 2, all accessions sharing the *MOT1*^{Sha} haplotype are both sensitive and unable to rescue the mutant phenotype.
(TIF)

Figure S6 Worldwide distribution of functionally contrasted alleles at *MOT1*. Original collection site and functional *MOT1* haplotype (Sha-like in red dots, L*er*-like in yellow, Col-like in blue) is shown on a world map for the 102 accessions sequenced in Table S1. L*er*-like accessions are found across the whole species known distribution range, while Sha-like accessions are restricted to Asia and Russia ('West-Asia').
(TIF)

Figure S7 Effect of Mo toxicity on fitness components -seed number and weight- of contrasted *MOT1* genotypes. The fitness consequences of Mo toxicity was tested on regular (non-acidic) soil mix with different nutrient solutions containing either traces of Mo ('Control') or 7 mM Na_2MoO_4 ('Na_2MoO_4 (7 mM)'). The assay was performed to compare (A) the *mot1.1* mutant and its wild-type ('WT') genetic background, (B) the Bay and Sha allele in the HIF084 background ('HIF[Bay]' vs 'HIF[Sha]'). In both cases, the defective *MOT1* allele is represented with black bars. To avoid heterogeneity/effects on descendance conveyed through the maternal plant, the mutant assays were performed as a progeny testing from a mother plant segregating for the T-DNA insertion. Two fitness parameters are represented: the total number of seeds produced per plant (on the left) and the weight of 1,000 seeds (on the right). Error bars show 95% confidence interval of the mean. For the weight of 1,000 seeds, there is no significant difference between genotypes under 'control' treatment, while defective *MOT1* alleles have significantly larger seeds under Mo excess (t-test; $p<0.017$ when comparing *mot1.1* and WT; $p<0.0018$ when comparing HIF084[Bay] and HIF084[Sha]).
(TIF)

Table S1 Haplotype diversity at MOT1 among 102 accessions. Polymorphisms detected across 102 *A. thaliana* accessions for the *MOT1* locus, including 1 kb upstream (promoter region) and 0.3 kb downstream of the coding region (between blue vertical double-lines). The position (coordinates) of polymorphic bases (regions) are indicated in bp from TAIR10 reference. Synonymous polymorphisms are highlighted in light grey, non-synonymous polymorphisms in medium grey and missing data in dark grey. *A. lyrata* and *A. halleri* serve as outgroups. Accessions individually phenotyped on acidic peatmoss substrate are indicated (S = Sensitive; R = Resistant). The main functional haplotypes highlighted with colours (Sha-like in red, L*er*-like in yellow, Col-like in blue) are those represented on Figure S6, including the Sha-like haplotype defined by the 'D104Y' polymorphism, and the Ler-like haplotype associated to the "53 bp-deletion."
(XLS)

Table S2 Genetic diversity and tests of selection at *MOT1* and two reference loci (*COI1* and *PI*).
(XLS)

Table S3 Detailed results of HKA simulations.
(XLS)

Acknowledgments

We thank I. Baxter, C. Toomajian, F. Roux, C. Camilleri, and F. Budar for discussions at various stages during this work and/or comments on the manuscript.

Author Contributions

Conceived and designed the experiments: OL SPK CT DES. Performed the experiments: SPK CT MA MS. Analyzed the data: SPK CT MS TR. Wrote the paper: OL SPK.

References

1. Karrenberg S, Widmer A (2008) Ecologically relevant genetic variation from a non-Arabidopsis perspective. Curr Opin Plant Biol 11: 156–162.
2. Nord EA, Lynch JP (2009) Plant phenology: a critical controller of soil resource acquisition. J Exp Bot 60: 1927–1937.
3. Baxter I, Brazelton JN, Yu D, Huang YS, Lahner B, et al. (2010) A coastal cline in sodium accumulation in *Arabidopsis thaliana* is driven by natural variation of the sodium transporter AtHKT1;1. PLoS Genet 6: e1001193. doi:10.1371/journal.pgen.1001193
4. Stinchcombe JR, Weinig C, Ungerer M, Olsen KM, Mays C, et al. (2004) A latitudinal cline in flowering time in *Arabidopsis thaliana* modulated by the flowering time gene *FRIGIDA*. Proc Natl Acad Sci USA 101: 4712–4717.
5. Trontin C, Tisné S, Bach L, Loudet O (2011) What does Arabidopsis natural variation teach us (and does not teach us) about adaptation in plants? Curr Opin Plant Biol 14: 225–231.
6. Hanikenne M, Talke IN, Haydon MJ, Lanz C, Nolte A, et al. (2008) Evolution of metal hyperaccumulation required *cis*-regulatory changes and triplication of *HMA4*. Nature 453: 391–395.
7. Turner TL, Bourne EC, Von Wettberg EJ, Hu TT, Nuzhdin SV (2010) Population resequencing reveals local adaptation of *Arabidopsis lyrata* to serpentine soils. Nat Genet 42: 260–263.
8. Alonso-Blanco C, Aarts MG, Bentsink L, Keurentjes JJ, Reymond M, et al. (2009) What has natural variation taught us about plant development, physiology, and adaptation? Plant Cell 21: 1877–1896.
9. Anderson JT, Willis JH, Mitchell-Olds T (2011) Evolutionary genetics of plant adaptation. Trends Genet 27: 258–266.
10. Hereford J (2009) A quantitative survey of local adaptation and fitness trade-offs. Am Nat 173: 579–588.
11. Fournier-Level A, Korte A, Cooper MD, Nordborg M, Schmitt J, et al. (2011) A map of local adaptation in *Arabidopsis thaliana*. Science 334: 86–89.
12. Hancock AM, Brachi B, Faure N, Horton MW, Jarymowycz LB, et al. (2011) Adaptation to climate across the *Arabidopsis thaliana* genome. Science 334: 83–86.
13. Baxter I, Muthukumar B, Park HC, Buchner P, Lahner B, et al. (2008) Variation in molybdenum content across broadly distributed populations of *Arabidopsis thaliana* is controlled by a mitochondrial molybdenum transporter (MOT1). PLoS Genet 4: e1000004. doi:10.1371/journal.pgen.1000004
14. Tomatsu H, Takano J, Takahashi H, Watanabe-Takahashi A, Shibagaki N, et al. (2007) An *Arabidopsis thaliana* high-affinity molybdate transporter required for efficient uptake of molybdate from soil. Proc Natl Acad Sci USA 104: 18807–18812.
15. Mengel K, Kirkby EA (2001) Principles of plant nutrition; Springer, editor. Berlin: Springer.
16. Mendel RR (2011) Cell biology of molybdenum in plants. Plant Cell Reports 30: 1787–1797.
17. Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, et al. (2010) Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature 467: 832–838.
18. Wood AR, Hernandez DG, Nalls MA, Yaghootkar H, Gibbs JR, et al. (2011) Allelic heterogeneity and more detailed analyses of known loci explain additional phenotypic variation and reveal complex patterns of association. Hum Mol Genet 20: 4082–4092.
19. Beck JB, Schmuths H, Schaal BA (2008) Native range genetic variation in *Arabidopsis thaliana* is strongly geographically structured and reflects Pleistocene glacial dynamics. Mol Ecol 17: 902–915.
20. Francois O, Blum MG, Jakobsson M, Rosenberg NA (2008) Demographic history of european populations of *Arabidopsis thaliana*. PLoS Genet 4: e1000075. doi:10.1371/journal.pgen.1000075
21. Platt A, Horton M, Huang YS, Li Y, Anastasio AE, et al. (2010) The scale of population structure in *Arabidopsis thaliana*. PLoS Genet 6: e1000843. doi:10.1371/journal.pgen.1000843
22. Nordborg M, Hu TT, Ishino Y, Jhaveri J, Toomajian C, et al. (2005) The pattern of polymorphism in *Arabidopsis thaliana*. PLoS Biol 3: e196. doi:10.1371/journal.pbio.0030196
23. Bergelson J, Roux F (2010) Towards identifying genes underlying ecologically relevant traits in *Arabidopsis thaliana*. Nat Rev Genet 11: 867–879.
24. Elwell AL, Gronwall DS, Miller ND, Spalding EP, Durham Brooks TL (2011) Separating parental environment from seed size effects on next generation growth and development in Arabidopsis. Plant Cell Environ 34: 291–301.
25. Krannitz PG, Aarssen LW, Dow JM (1991) The effect of genetically based differences in seed size on seedling survival in *Arabidopsis thaliana* (Brassicaceae). Am J Bot 78: 446–450.
26. Kroymann J, Donnerhacke S, Schnabelrauch D, Mitchell-Olds T (2003) Evolutionary dynamics of an Arabidopsis insect resistance quantitative trait locus. Proc Natl Acad Sci USA 100: 14587–14592.
27. Todesco M, Balasubramanian S, Hu TT, Traw MB, Horton M, et al. (2010) Natural allelic variation underlying a major fitness trade-off in *Arabidopsis thaliana*. Nature 465: 632–636.
28. Zhen Y, Dhakal P, Ungerer MC (2011) Fitness benefits and costs of cold acclimation in *Arabidopsis thaliana*. Am Nat 178: 44–52.
29. Loudet O, Chaillou S, Camilleri C, Bouchez D, Daniel-Vedele F (2002) Bay-0×Shahdara recombinant inbred line population: a powerful tool for the genetic dissection of complex traits in Arabidopsis. Theor Appl Genet 104: 1173–1184.
30. Loudet O, Gaudon V, Trubuil A, Daniel-Vedele F (2005) Quantitative trait loci controlling root growth and architecture in *Arabidopsis thaliana* confirmed by heterogeneous inbred family. Theor Appl Genet 110: 742–753.
31. Kronholm I, Loudet O, de Meaux J (2010) Influence of mutation rate on estimators of genetic differentiation–lessons from *Arabidopsis thaliana*. BMC Genet 11: 33.
32. Soltanpour PP, Schwab AP (1977) A new soil test for simultaneous extraction of macro- and micronutrients in alkaline soils. Comm Soil Sci Plant Anal 8: 195–207.
33. McKhann HI, Camilleri C, Berard A, Bataillon T, David JL, et al. (2004) Nested core collections maximizing genetic diversity in *Arabidopsis thaliana*. Plant J 38: 193–202.
34. Gode C, Decombeix I, Kostecka A, Wasowicz P, Pauwels M, et al. (2012) Nuclear microsatellite loci for *Arabidopsis halleri* (Brassicaceae), a model species to study plant adaptation to heavy metals. Am J Bot.
35. Caldwell KS, Michelmore RW (2009) *Arabidopsis thaliana* genes encoding defense signaling and recognition proteins exhibit contrasting evolutionary dynamics. Genetics 181: 671–684.
36. Cork JM, Purugganan MD (2005) High-diversity genes in the Arabidopsis genome. Genetics 170: 1897–1911.
37. Nei M (1987) Molecular Evolutionary Genetics; Press CU, editor. New York: Columbia Univ. Press.
38. Watterson GA (1975) On the number of segregating sites in genetical models without recombination. Theor Popul Biol 7: 256–276.
39. Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585–595.
40. McDonald JH, Kreitman M (1991) Adaptive protein evolution at the Adh locus in Drosophila. Nature 351: 652–654.
41. Rand DM, Kann LM (1996) Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from Drosophila, mice, and humans. Mol Biol Evol 13: 735–748.

Supplementary Materials:

- Supplementary Tables

Table S1: Haplotype diversity at MOT1 among 102 accessions

Polymorphisms detected across 102 *A. thaliana* accessions for the *MOT1* locus, including 1kb upstream (promoter region) and 0.3kb downstream of the coding region (between blue vertical double-lines). The position (coordinates) of polymorphic bases (regions) are indicated in bp from TAIR10 reference. Synonymous polymorphisms are highlighted in light grey, non-synonymous polymorphisms in medium grey and missing data in dark grey. *A. lyrata* and *A. halleri* serve as outgroups. Accessions individually phenotyped on acidic peatmoss substrate are indicated (S = Sensitive; R = Resistant). The main functional haplotypes highlighted with colours (Sha-like in red, L*er*-like in yellow, Col-like in blue) are those represented on Figure S6, including the Sha-like haplotype defined by the 'D104Y' polymorphism, and the Ler-like haplotype associated to the '53bp-deletion'.

Table S2: Genetic diversity and tests of selection at *MOT1* and two reference loci (*COI1* and *PI*)

Table S3: Detailed results of HKA simulations

- Supplementary Figures

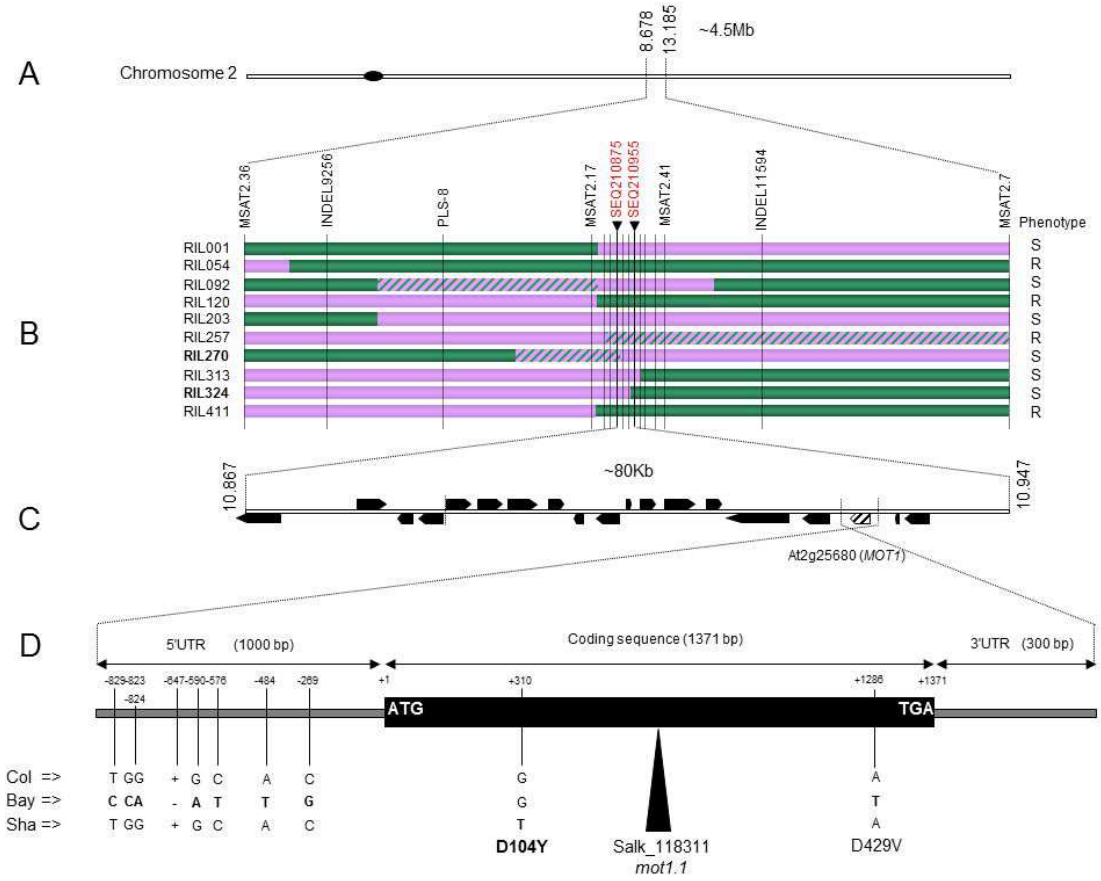


Figure S1: Fine-mapping the causative locus identifies *MOT1* as a candidate gene

A. The physical region of chromosome 2 found to be linked to the growth defect phenotype is shown (physical positions are given in Mb). B. Zooming in the candidate region highlights recombinants within the inbred lines that allow to fine-map the causative locus, thanks to additional markers (vertical dashed lines). Individual lines' genotype are depicted in horizontal coloured boxes (green = Bay allele; purple = Sha allele ; dashed = heterozygous) and their phenotype on peatmoss are indicated (S = Sensitive; R = Resistant). C. This allows to narrow down the causative region to 80kb, a region containing 19 predicted genes including the candidate At2g25680 (*MOLYBDATE TRANSPORTER 1*). D. *MOT1* has been sequenced in parental accessions and polymorphisms between Col-0, Bay-0 and Shahdara are represented along the single-exon gene model, including an amino-acid change specific to Shahdara (D104Y). The position of the insertion of a T-DNA in the *mot1.1* mutant (SALK_118311) is indicated.

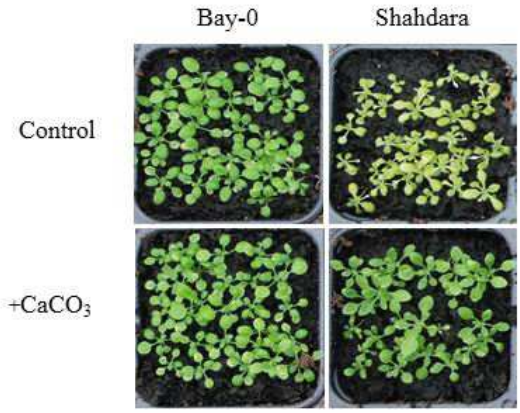


Figure S2: Chemical complementation links growth defect with soil pH

Increasing soil pH from ~5 ('Control') to ~6 ('+$CaCO_3$') by doubling the amount of $CaCO_3$ mixed to the peatmoss substrate rescues normal vegetative growth of the Shahdara strain.

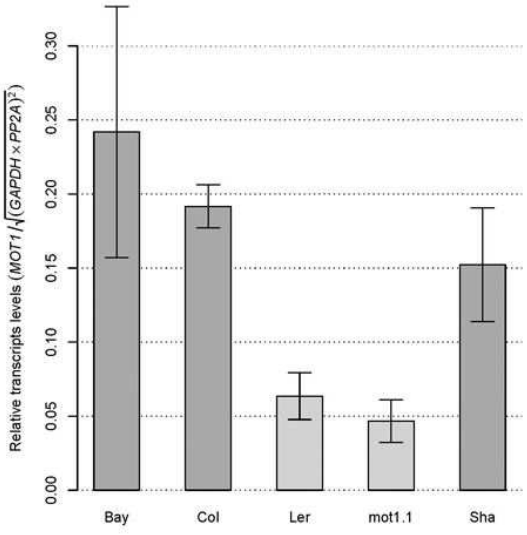


Figure S3: *MOT1* transcript accumulation does not explain *MOT1*^{Sha} defective allele

MOT1 transcript accumulation relative to *GAPDH* and *PP2A* controls is shown from roots of diverse genotypes as in Figure 3. Contrary to L*er* and *mot1.1*, Shahdara accumulates normal levels of transcript. Standard errors are shown.

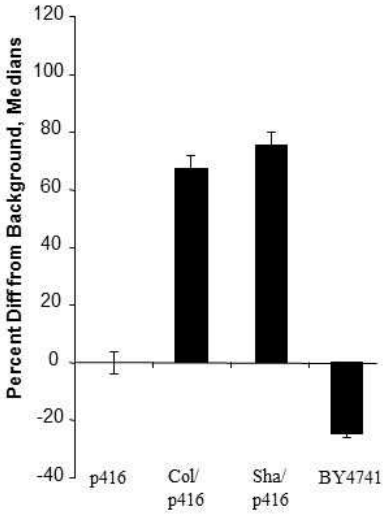


Figure S4: MOT1^{Sha} is able to transport Mo in yeast

The Sha allele of MOT1 was overexpressed in yeast heterologous system (Sha/p416) and shown to lead to Mo accumulation compared to the empty vector (p416, used as reference) or the yeast wild-type strain (BY4741), to an extent not significantly different from the Col MOT1 allele (Col/p416). Error bars represent interquantile range of medians.

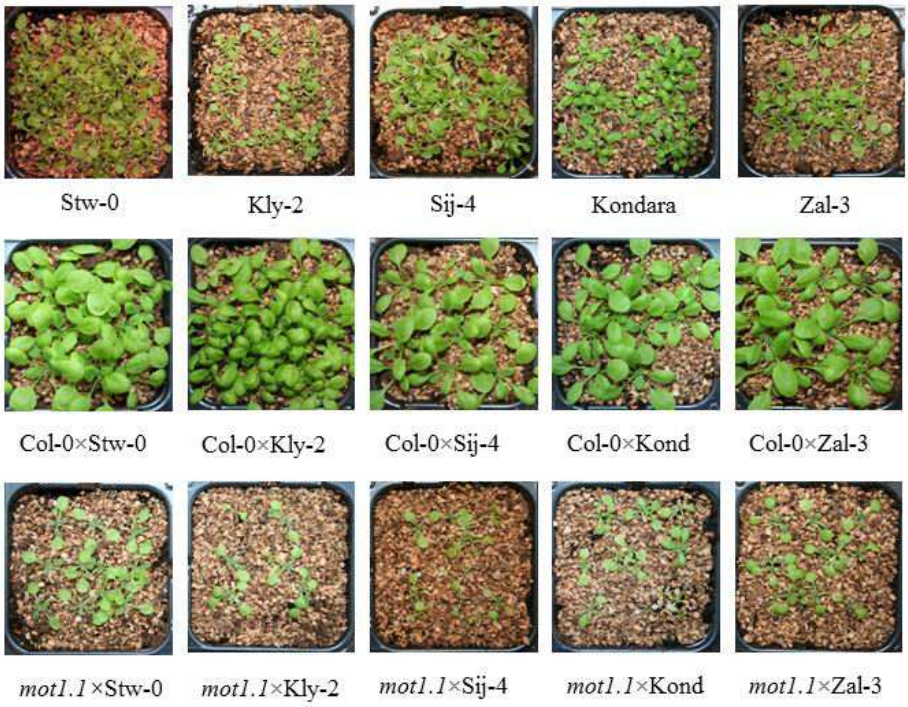


Figure S5: Multiple accessions sharing the *MOT1*^{Sha} haplotype confirm the causative gene and polymorphism

Peatmoss phenotype of diverse genotypes is shown: 5 independent Sha-like accessions (Stw-0, Kly-2, Sij-4, Kondara and Zal-3) and F1 plants from complementation crosses between each of these accessions and either the *mot1.1* mutant or its wild-type genetic background (Col-0). As in Figure 2, all accessions sharing the *MOT1*^{Sha} haplotype are both sensitive and unable to rescue the mutant phenotype.

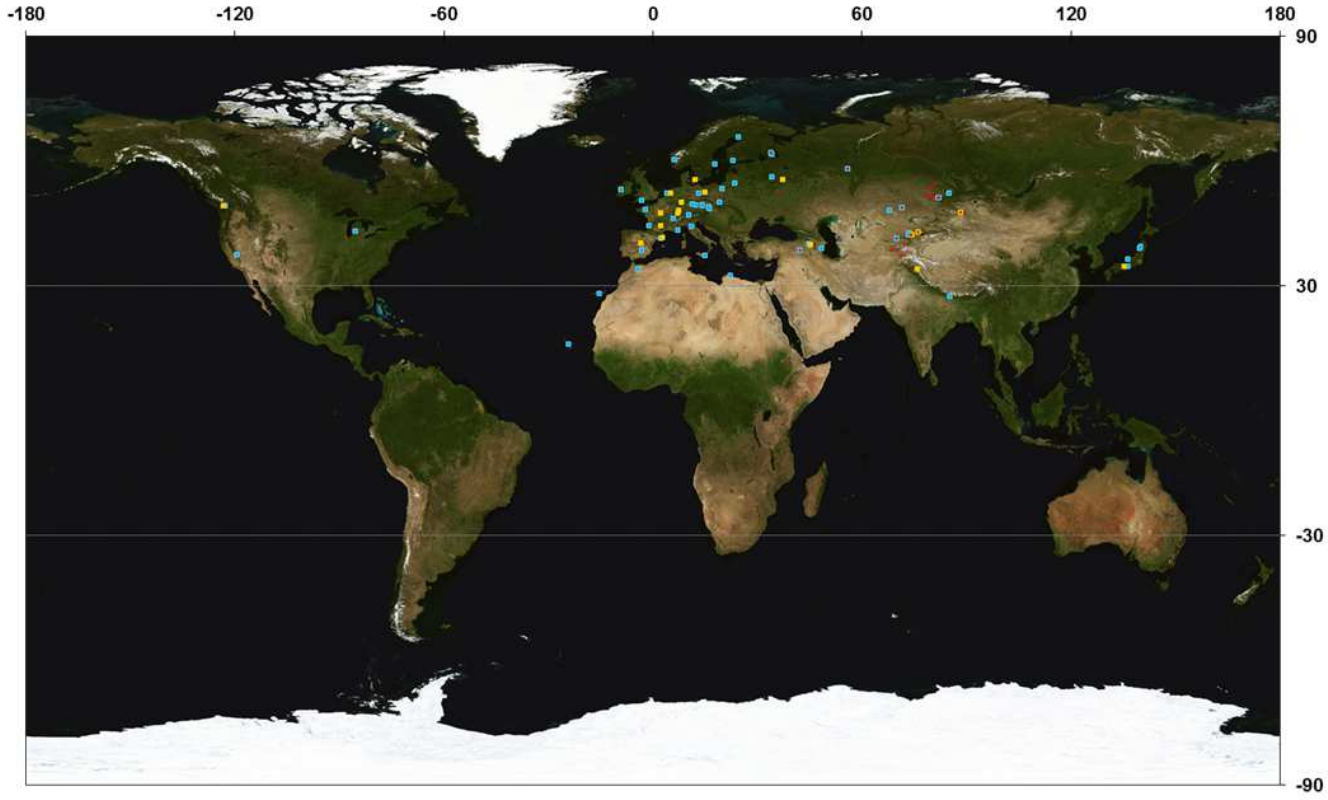


Figure S6: Worldwide distribution of functionally contrasted alleles at *MOT1*

Original collection site and functional *MOT1* haplotype (Sha-like in red dots, L*er*-like in yellow, Col-like in blue) is shown on a world map for the 102 accessions sequenced in Table S1. L*er*-like accessions are found across the whole species known distribution range, while Sha-like accessions are restricted to Asia and Russia ('West-Asia').

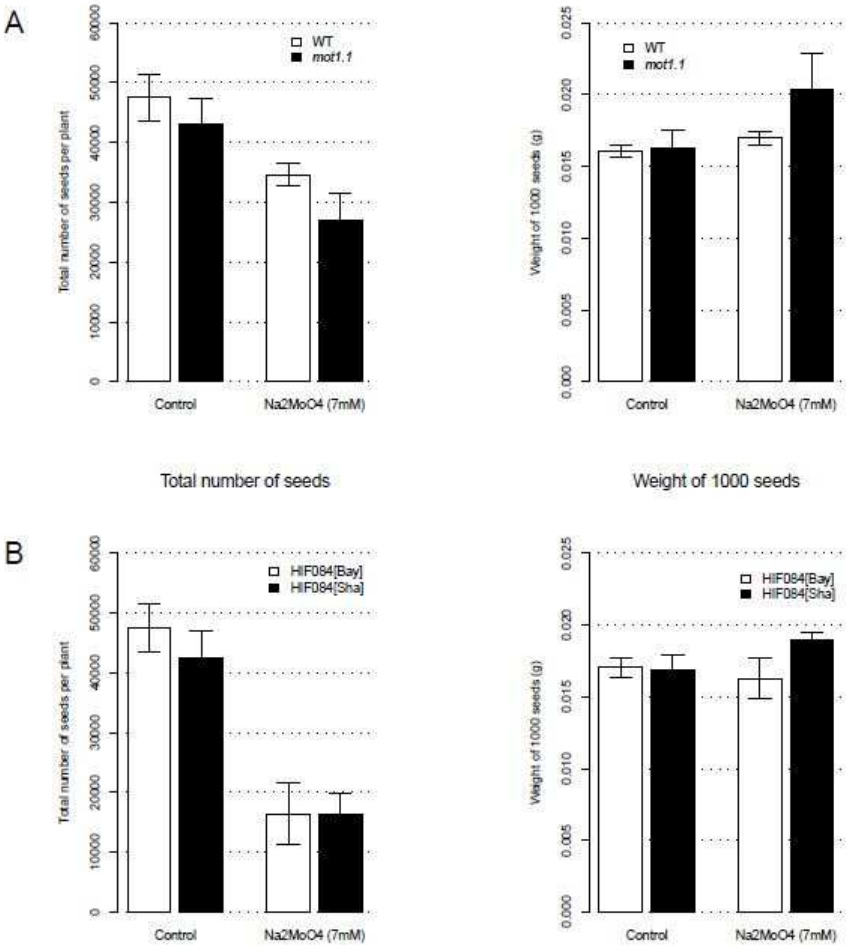
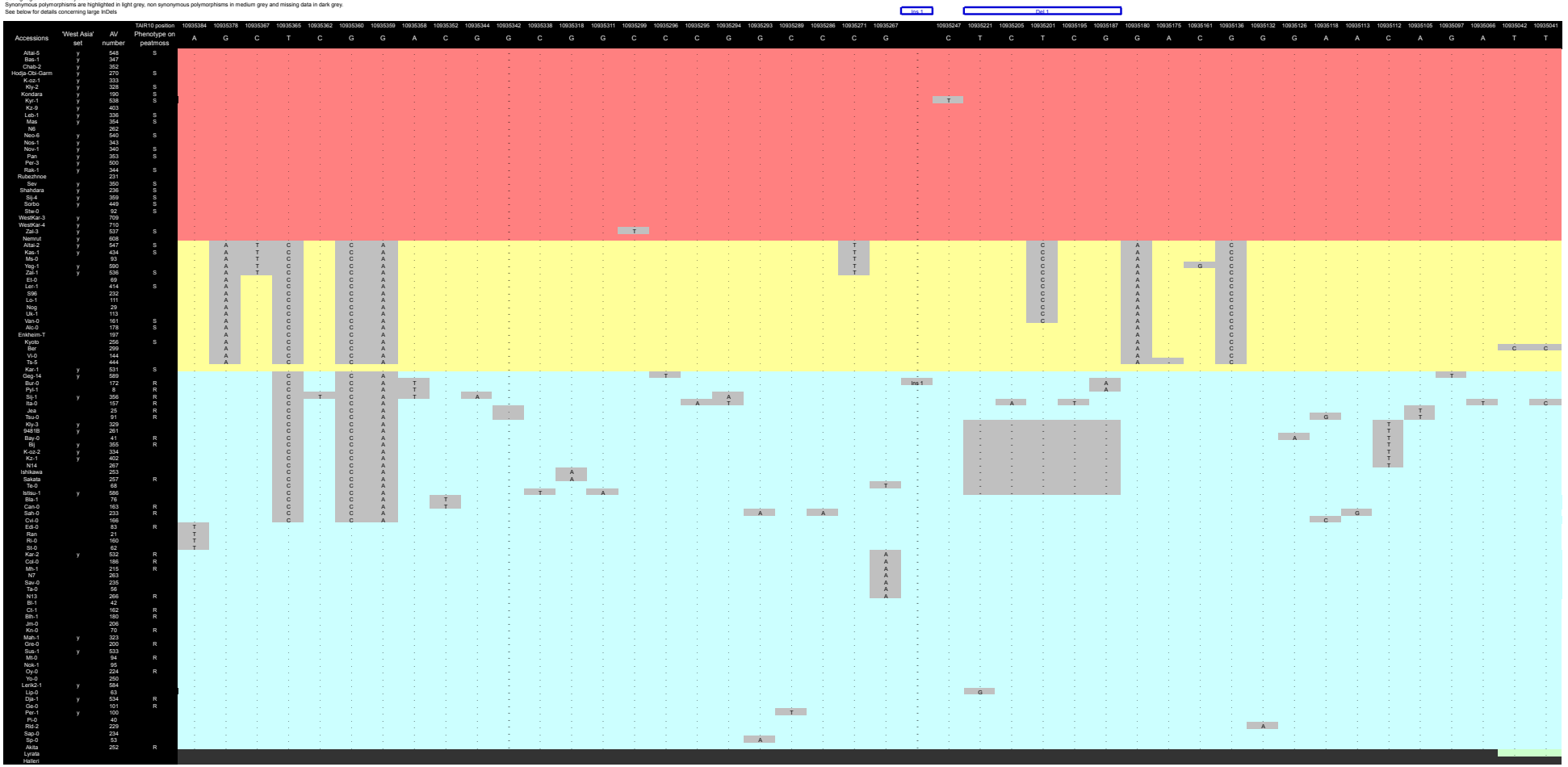


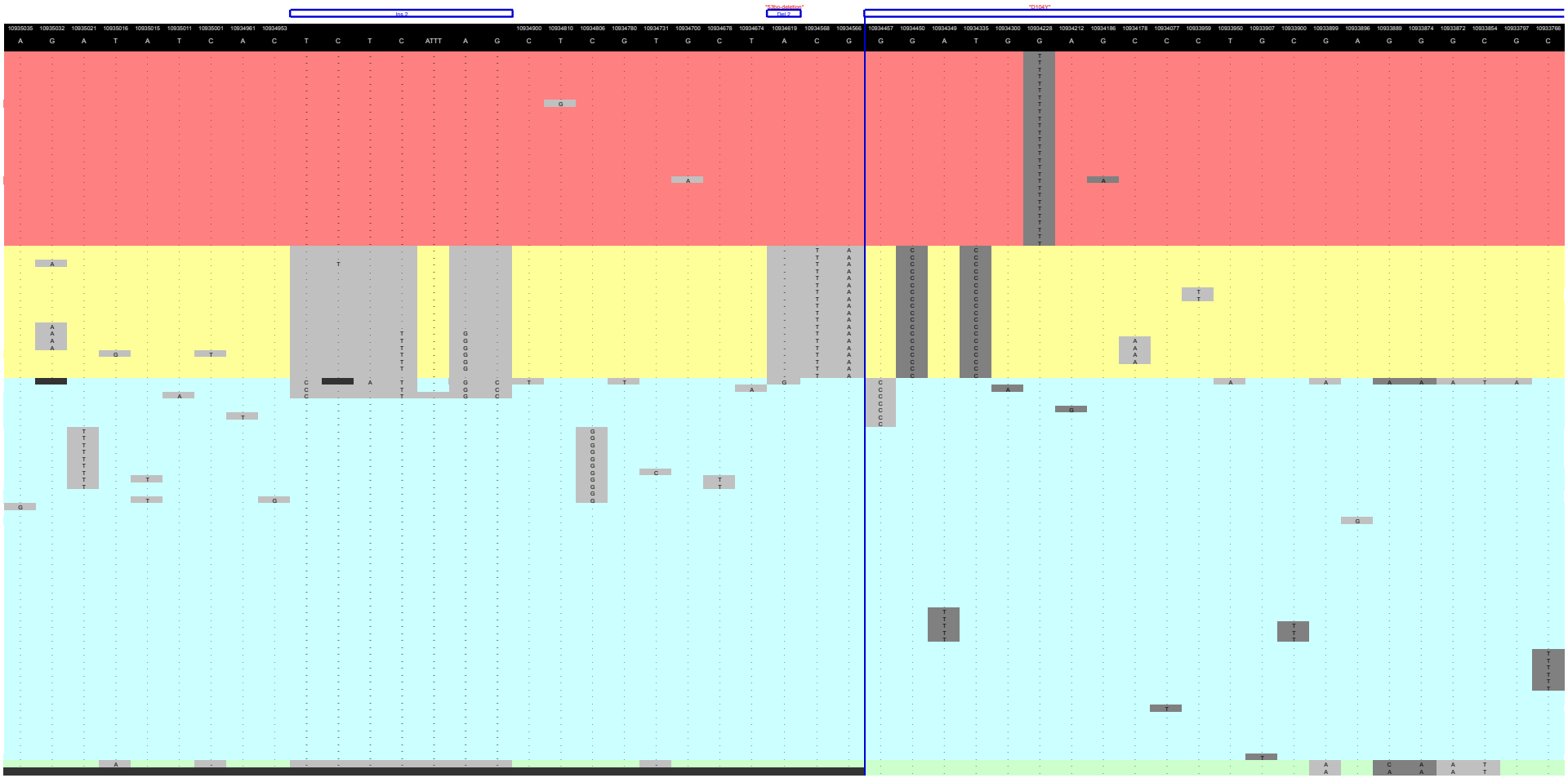
Figure S7: Effect of Mo toxicity on fitness components -seed number and weight- of contrasted *MOT1* genotypes

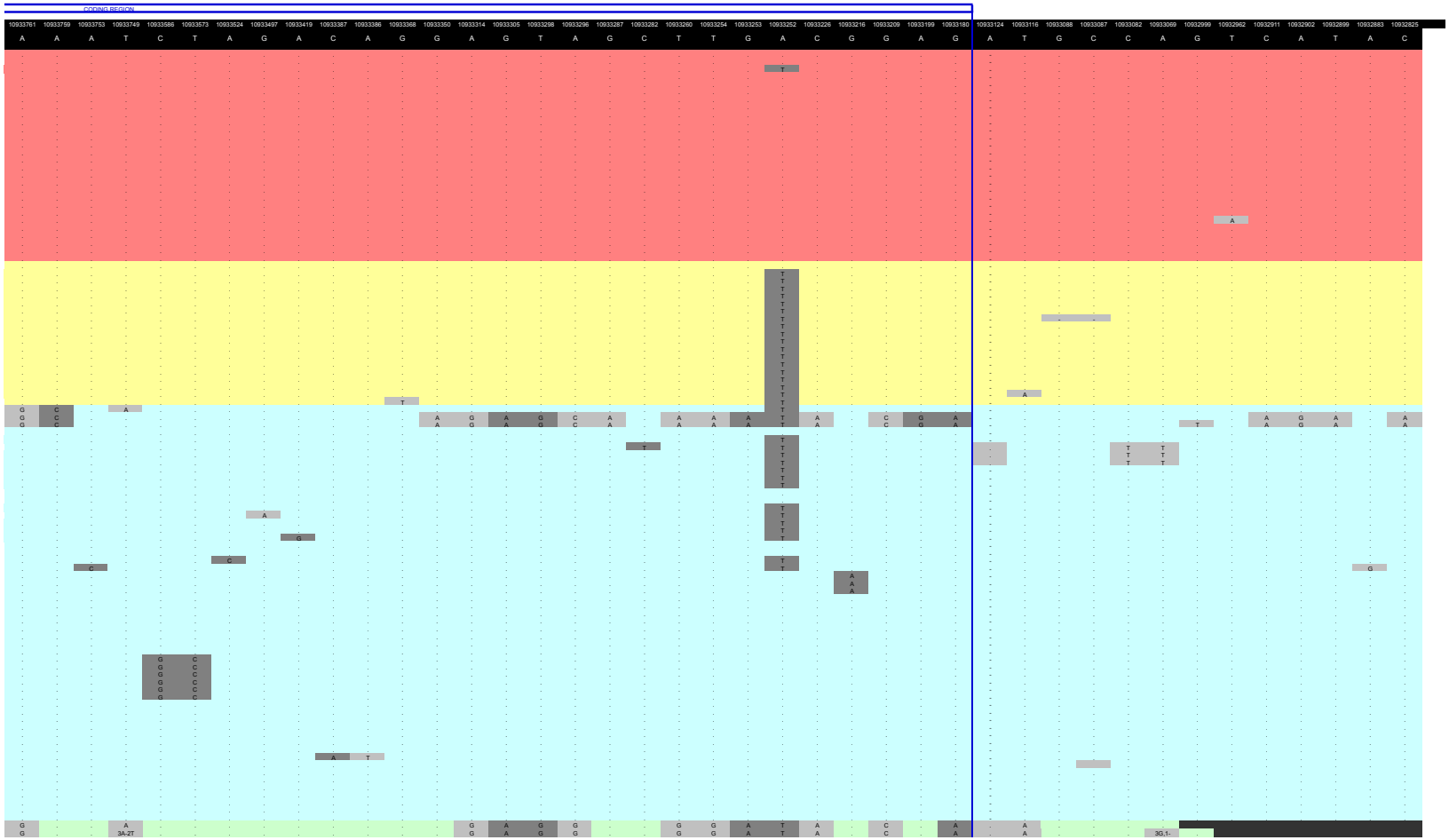
The fitness consequences of Mo toxicity was tested on regular (non-acidic) soil mix with different nutrient solutions containing either traces of Mo ('Control') or 7mM Na_2MoO_4 ('Na_2MoO_4 (7mM)'). The assay was performed to compare (A) the *mot1.1* mutant and its wild-type ('WT') genetic background, (B) the Bay and Sha allele in the HIF084 background ('HIF[Bay]' vs 'HIF[Sha]'). In both cases, the defective *MOT1* allele is represented with black bars. To avoid heterogeneity / effects on descendance conveyed through the maternal plant, the mutant assays were performed as a progeny testing from a mother plant segregating for the T-DNA insertion. Two fitness parameters are represented: the total number of seeds produced per plant (on the left) and the weight of 1,000 seeds (on the right). Error bars show 95% confidence interval of the mean. For the weight of 1,000 seeds, there is no significant difference between genotypes under 'control' treatment, while defective *MOT1* alleles have significantly larger seeds under Mo excess (t-test; $p<0.017$ when comparing *mot1.1* and WT; $p<0.0018$ when comparing HIF084[Bay] and HIF084[Sha]).

Supplementary Table 1: Haplotype diversity at *MOT1* among 102 accessions.
Sha-like accessions are highlighted in red, Ler-like in yellow and Col-like in blue.
Coding region is delimited by blue vertical lines.
Synonymous polymorphisms are highlighted in light grey, non synonymous polymorphisms in medium grey and missing data in dark grey.
See below for details concerning large InDels



Ins 1: insertion of 322bp between position 10935257 and 10935258. The sequence of the Bur-0 insertion is given below.
TCGAGCAAACTAGAAAAAGAGATCGACGCATCATCTCTAGGATCAATCGATCTCAACATCTATACTTGAGAATTTTAAAAATATTTCAACTTTGATGCTGATTCATCTCATCTTTTTTAGAGGATTTTCACATTTATAAAAGGCTTATTGGATGCCTAACTGATTCTATCCGAAATTCATTCGGTTCCTATAAAGTAAAAACGGATCTTAAACTCTCAACCCATAATCTCAAAATTTGAATTACTCAATATAAAATAATTCGTGAAACCCGAATGTTCAAGACTACCGGAAACATCTGTGAAATATAAAGCTTTGTTTGTTAAT
Del 1: Deletion of 49bo from 10935231 to 10935183. The sequence of the deletion is given below.
ATCACAAAAATTAATAAACTAAATTTCTATTATATGCTTAATGGGTTTT
Ins 2: Insertion of 322bp (326bp for Pyl-1) between position 10934952 and 10934953. The sequence of Pyl-1 is given below. Polymorphisms within the insertion are highlighted in bold.
C**A**GCTCT**A**TTTGTATATAAGTGTTGACTTTTGATATATATAAATACCTTTCCATAAATGAATAATTTATGTTGCTTAGATTTATATAAATAACCTACAAATTCAGGTGTTAAATGTTTAACTGAAGATATAATAAAACAATAGAGATAGGAAATGATACTTAGTCTTACAAATTTAACTAATAGTACAAACAATA**T**TATTTAAAACAAAAGTAAAACTTGAAAACAACTCCGTAAAACAGAATAAATA**T**AAAATGAATTTATTTATTTATTT**ATTT**GATGGCGACTTTTCTATTGCCG**G**TTCAAAACTCAAATAATAAATAC**A**CA
Del 2: Deletion of 53bp from 10934626 to 10934574. The sequence of the deletion is given below.
TTGCTTTATTGTCAACGCTTTGGTTTTTGATACGATATAAAGAGATACGCTTA


Ins 2
"Elba-deletion"
Del 2
"D1041"
10935035 10935032 10935021 10935016 10935015 10935011 10935001 10934961 10934953
10934900 10934810 10934806 10934780 10934731 10934700 10934678 10934674 10934619 10934588 10934566
10934457 10934450 10934349 10934335 10934300 10934228 10934212 10934186 10934178 10934077 10933959 10933950 10933907 10933900 10933899 10933896 10933889 10933874 10933872 10933854 10933797 10933766


CODING REGION

Supplementary Table 2: Genetic diversity and tests of selection at *MOT1* and two reference loci (*COI1* and *PI*).

		N (a)	Length	S (b)	Singletons	π (c)	θw (d)	Tajima's D (e)	MK test (f)	HKA
MOT1	Worldwide	102	2444 (g)	114 (117)	57	0.00348	0.00897	**-2,059 (0,018*)**	-	-
		102	1428 (h)	51	22	0.00234	0.00687	-	**3,02 (0,036*)**	**12.445 (0,014*) (i)**
	West-Asia	44	2447 (g)	55	37	0.00252	0.00517	**-1,806 (0,014*)**	-	-
		44	1428 (h)	21	16	0.00146	0.00338	-	1,615 (0,55)	9.121 (0,058) (i)
	A. halleri	5	1428 (h)	36	21	0.01218	0.0121	0,0521 (0,58)	-	-
COI1	Worldwide	102	2513 (g)	33	13	0.00181	0.00253	-0,863 (0,21)	-	-
		102	2286 (h)	29	13	0.00154	0.00244	-	0,51 (0,684)	1.639 (0,44) (j)
	West-Asia	44	2551 (g)	22	6	0.00197	0.00198	-0,0215 (0,55)	-	-
		44	2322 (h)	17	4	0.00163	0.00168	-	0,6875 (1)	1.097 (0,58) (j)
	A. halleri	5	2322 (g)	19	13	0.00353	0.00393	-0,744 (0,31)	-	-
PI	Worldwide	102	2098 (g)	46	19	0,00230	0.00422	-1,432 (0,051)	-	-
		102	1313 (h)	33	15	0.00245	0.00484	-	NA	1.639 (0,44) (j)
	West-Asia	44	2105 (g)	17	7	0.00119	0.00186	-1,137 (0,121)	-	-
		44	1317 (h)	10	7	0.00109	0.00175	-	NA	1.097 (0,58) (j)
	A. halleri	5	1317 (h)	29	20	0.01002	0.01057	-0,386 (0,44)	-	-

(a) Number of sequences
(b) Number of segregating sites in the region (number of mutations if different from S)
(c) Number of nucleotide differences per site between two sequences
(d) Theta per site from S
(e) Tajima's D statistics value (p-value)
(f) McDonald-Kreitman test using substitutions in coding regions without singletons. Neutral index (p-value of Fisher statistics)
(g) Length of the sequenced region (in bp)
(h) Length of the sequenced region common between A. thaliana and A. halleri (in bp)
(i) Multilocus HKA test was performed using MOT1, PI and COI loci. Sum of deviation (probability from Chi square distribution)
(j) Multilocus HKA test was performed using PI and COI loci. Sum of deviation (Probability from Chi-square distribution)
* Significant after modified Bonferroni correction
NA The statistics could not be estimated due to a lack of non-synonymous polymorphism

Supplementary Table 3: Detailed results of HKA simulations

			Polymorphic sites within species		Divergence between species	
Sample	**Locus**		**Observed (a)**	**Expected (b)**	**Observed (c)**	**Expected (d)**
Worldwide	MOT1	A. thaliana	51	32.88	79.53	108.62
		A. halleri	36	25.02		
	COI	A. thaliana	29	42.43	165.90	140.17
		A. halleri	20	32.99		
	PI	A. thaliana	33	37.69	127.90	124.52
		A. halleri	30	28.69		
West-Asia	MOT1	A. thaliana	21	12.95	79.48	100.87
		A. halleri	36	22.66		
	COI	A. thaliana	17	19.11	165.40	148.84
		A. halleri	19	33.44		
	PI	A. thaliana	10	15.94	128.95	124.12
		A. halleri	29	27.89		

(a) Number of segregating sites observed at MOT1 and the two reference loci in *A. thaliana* and *A. halleri*.

(b) Number of segregating sites expected at MOT1 and the two reference loci in A. thaliana and A. halleri after 10.000 simulations performed using HKA multilocus software.

(c) Average number of nucleotide differences observed between population at MOT1 and two reference loci.

(d) Average number of nucleotide differences expected between population at MOT1 and two reference loci after 10.000 simulations performed using HKA multilocus software.

7. GENERAL DISCUSSION AND PERSPECTIVES

7.1 Why is MOT1[Sha] hypofunctional?

In this study, we identified a new variant of the MOT1 transporter in the Shahdara accession, characterized by the D104Y mutation. From a phenotypic point of view, everything converges to say that *MOT*1[*Sha*] is hypofunctional. All tested accessions carrying the *MOT1[Sha]* allele presented a reduced shoot growth phenotype on peatmoss that was complemented by increasing pH and soil Mo concentration, i.e. by increasing Mo availability (Publication – Fig. 3, Fig.S2). They also presented low level of Mo both in seeds and in leaves according to the ionomics database. Nevertheless, the reason for this hypofunctionality remains unknown. We showed in yeast heterologous system that MOT1[Sha] results in Mo accumulation at the same level as MOT1[Col]. It is possible that this system is not sensitive enough to detect fine functional differences between the two alleles, especially when overexpressed. Besides, some data indicate that in yeast heterologous system, like in plant cells, MOT1 transporter is targeted to the mitochondria. As a consequence, the accumulation of Mo in yeast might not result from direct transport through MOT1 but from a complicated, indirect and unknown mechanism. Patch clamp experiments could have been necessary to detect subtle differences in MOT1[Col] and MOT1[Sha] functionality.

In *MOT2*, two mutations in the signal peptide of the protein have been shown to interfer with MOT2 localisation at the vacuolar membrane [277]. In our case, it is unlikely that the D104Y mutation would affect MOT1 mitochondrial localisation because the polymorphism is not located in the signal peptide and localisation prediction software predicts MOT1[Sha] at the mitochondria like MOT1[Col] (TargetP). However, we could have done MOT1[Sha]-GFP fusions to confirm properly that MOT1[Sha] is targeted to the mitochondria.

The aspartic acid at position 104 in MOT1 is located in a short region that is not well conserved within the homologs of MOT1 and corresponds to a histidine in MOT2 homologs. This relatively poor conservation might suggest that this amino acid is not functionally important for Mo transport (figure 23) (which is consistent with results in yeast heterologous system). Nevertheless, it was striking to observe that the mutation of this amino acid could modify transmembrane predictions (TMHMM predictions v2, figure 24). Thus, D104Y could somehow destabilize the transporter or prevent its interaction with other proteins. Overall, the poor

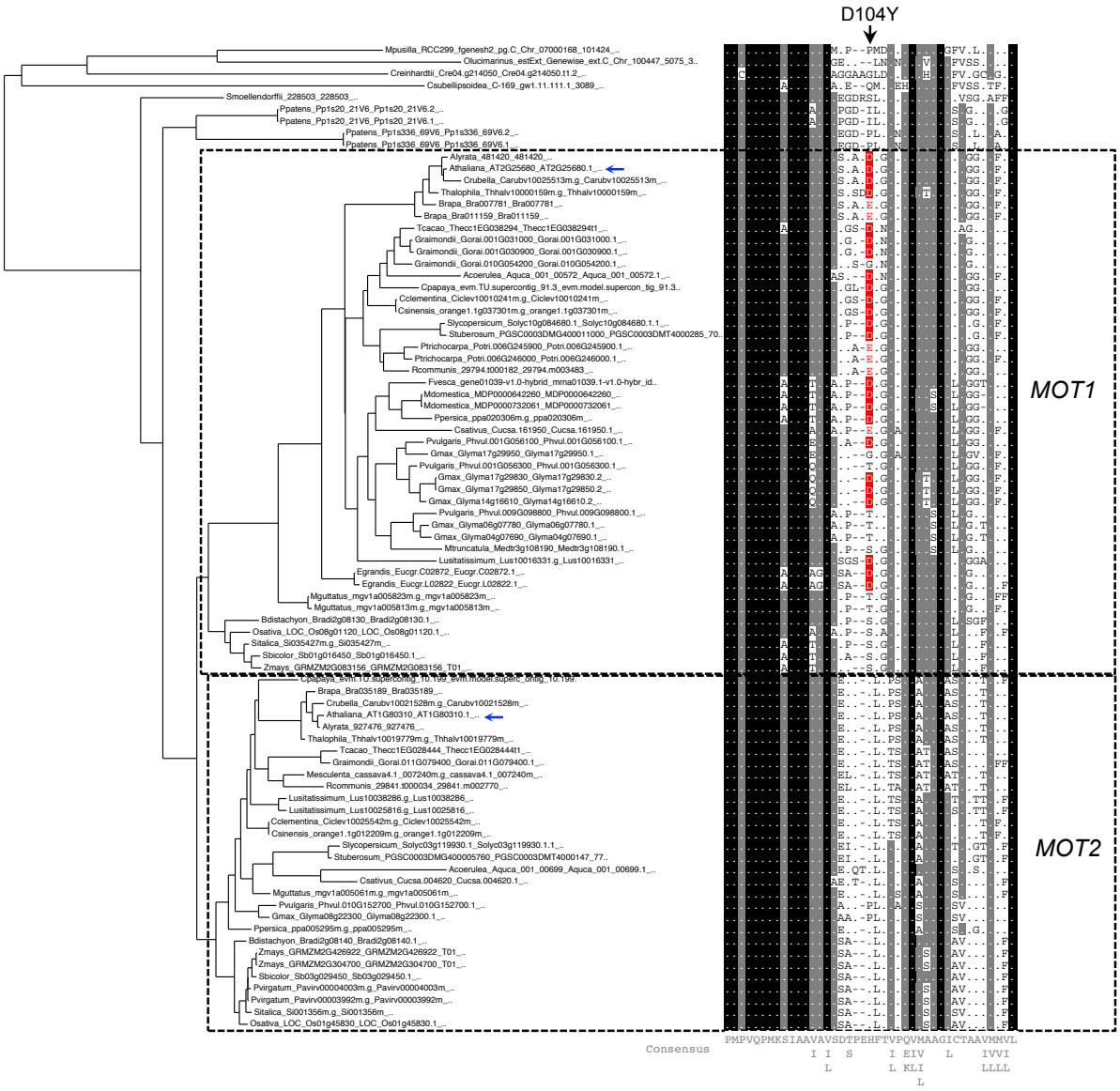


Fig. 23. Phylogenetic analysis of *MOT1* and *MOT2* homologs retrieved from PhytozomeV9 for Fabid, Malvid, Grass and Chlorophyte (but see below). The name of each sequence is composed of the species name followed by the locus name and transcript name as retrieved from Phytozome website. The sequences encompassing the 62 amino acids before and the 99 amino acids after D104 amino acid (*AtMOT1*) were aligned using Muscle implemented in SeaviewV4. Phylogeny analysis was performed on DNA sequences using PhyML implemented on www.phylogeny.fr website. *AtMOT1* and *AtMOT2* are indicated by blue arrows. The proteic alignment of the 20 amino acids before and after D104 amino acid (AtMOT1) is indicated. The dot correspond to the functionally conserved amino acids indicated on the consensus. Note that D104 amino acid in AtMOT1 (or functionally equivalent glutamic acid) is not conserved within the MOT1 clade and correspond to an histidine in the MOT2 clade. Some sequences were not included in the phylogeny because they presented too many deletions in the interesting region (Vvinifera|GSVIVG01035740001|GSVIVT01035740001; Egrandis|Eucgr.J03028|Eucgr.J03028.1; Egrandis|Eucgr.C02871|Eucgr.C02871.1; Egrandis|Eucgr.A01699|Eucgr.A01699.1; Egrandis|Eucgr.A01696|Eucgr.A01696.1; Egrandis|Eucgr.A01695|Eucgr.A01695.1; Egrandis|Eucgr.A01694|Eucgr.A01694.1; Egrandis|Eucgr.A01693|Eucgr.A01693.1).

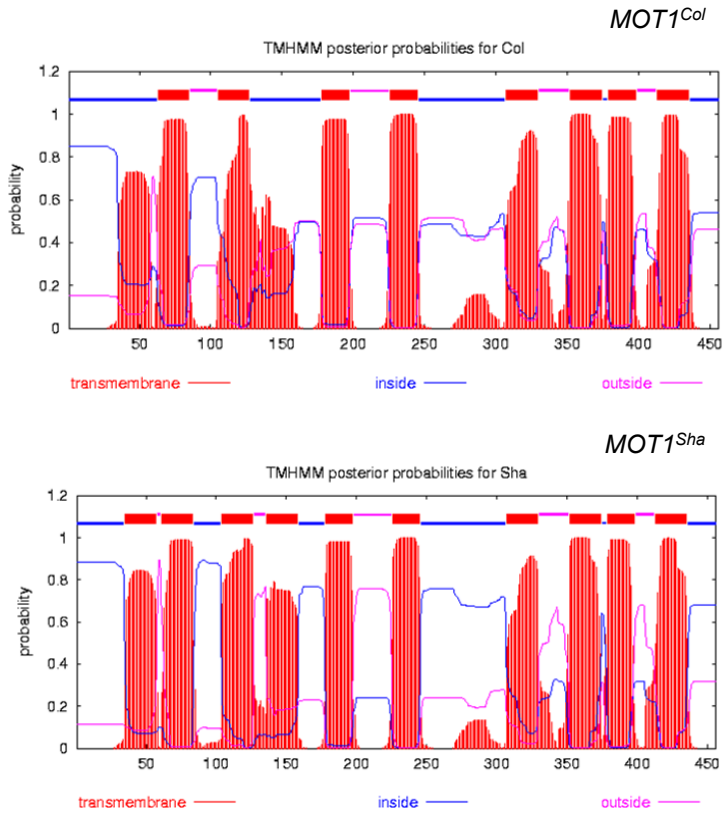


Fig. 24. Transmembrane predictions of MOT1[Col] and MOT1[Sha] transporters using TMHMM software. Note that the D104Y mutation in Sha altered transmembrane predictions around amino acids 130.

understanding of MOT1 function does not help understanding MOT1[Sha] hypofunctionality.

7.2 Is MOT1[Sha] adaptive?

Another conclusion from our paper concerns the possible local selection of the *MOT1[Sha]* allele in 'West Asia' in response to soils rich in Mo. This conclusion is supported by two facts: *MOT1[Sha]* worldwide distribution together with the pattern of polymorphisms; the striking occurrence of *MOT1[Sha]* on soils rich in Mo.

Regarding the first point, the low levels of polymorphism associated with the excess of rare variants (Tajima's D<0) observed at *MOT1* in Sha-like accessions are consistent with the pattern expected under strong positive selection and the absence of such patterns in the neutral genes suggest that it might not be due to demographic events (table 4). Nevertheless, these patterns are particularly difficult to interpret in *A. thaliana*. Indeed, an excess of rare polymorphisms has been observed in that species resulting in a distribution of Tajima's D

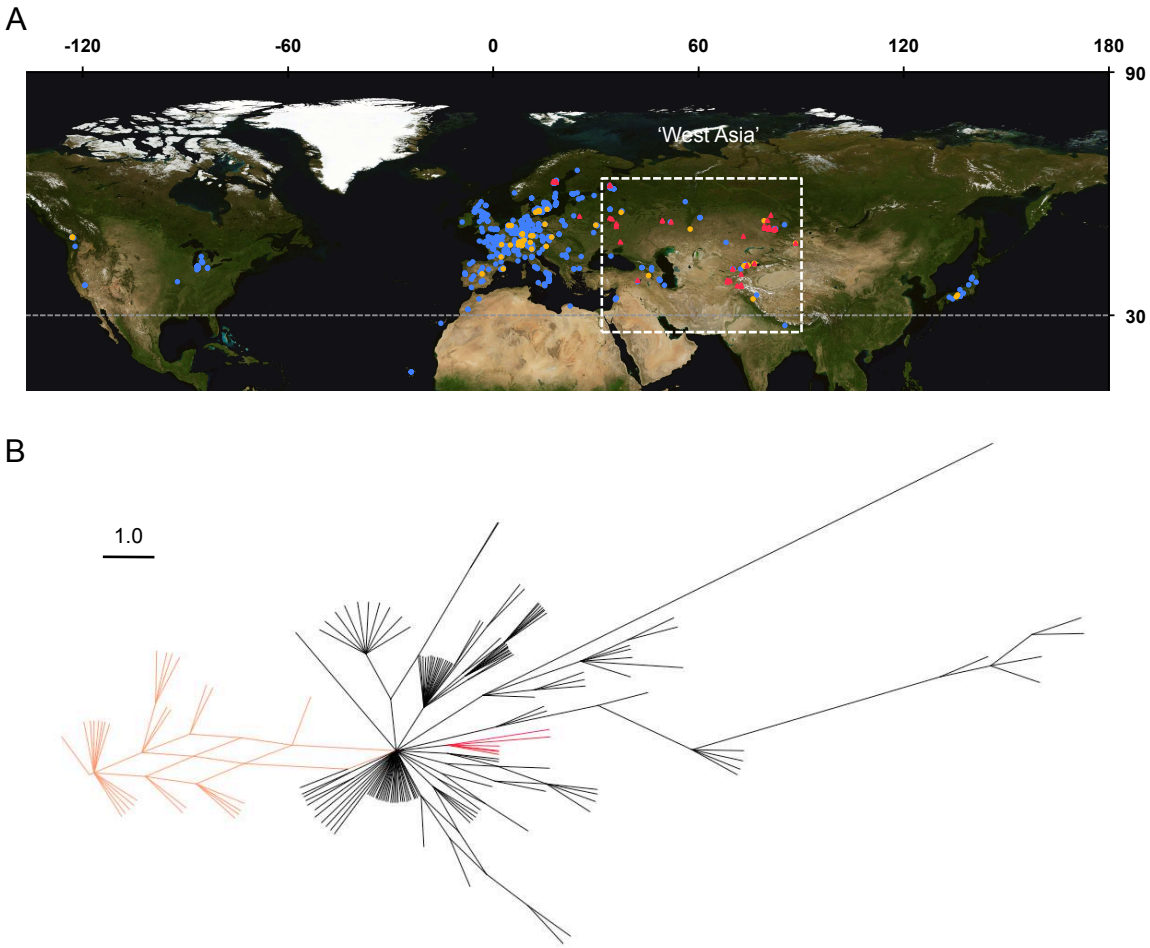


Fig. 25. *MOT1* natural variation in *A. thaliana*. A. The original collection site of Sha-like, Ler-like and Col-like accessions (including newly sequenced 1001 genomes accessions) have been localised on a worldwide map. Over 600 accessions, 53 Sha-like and 43 Ler-like acessions were identified. The region denominated as 'West Asia'in the text is highlighted. B. An haplotype network of MOT1 using over 500 accessions have been generated using Splits tree V4 and median network analysis. Ler-like accessions (orange) and Sha-like (red) are grouped together

Tab. 4. Genetic diversity at MOT1 and two reference loci (COI1 and PI) in Sha-like, Van-like and Col-like subgroups.

			N (a)	Length (b)	S (c)	Singletons	π (d)	Tajima's D (e)
MOT1	**Sha-like**	Worldwide	28	2576	7	7	0.00019	-2.169 (0.002**)
		Asiatic	25	2576	5	5	0.00016	-1.999 (0.004**)
	Van-like	Worldwide	19	2841	23	15	0.00168	-1.076 (0.14)
		Asiatic	5	2844	12	12	0.00169	-1.205 (0.041*)
	Col-like	Worldwide	55	2519	91 (94)	47	0.00331	-2.094 (0.003**)
		Asiatic	14	2500	40	32	0.00337	-1.434 (0.064)
COI1	**Sha-like**	Worldwide	28	2537	17	3	0.0021	0.757 (0.82)
		Asiatic	25	2551	17	3	0.00197	0.403 (0.71)
	Van-like	Worldwide	19	2537	12	2	0.00169	0.898 (0.84)
		Asiatic	5	2562	9	1	0.00203	1.448 (0.93)
	Col-like	Worldwide	55	2526	30	12	0.00168	-1.154 (0.11)
		Asiatic	14	2738	19	9	0.00222	-0.097 (0.50)
PI	**Sha-like**	Worldwide	28	2115	11	5	0.00114	-0.472 (0.35)
		Asiatic	25	2117	7	2	0.00093	0.188 (0.63)
	Van-like	Worldwide	19	2126	24	11	0.00282	-0.492 (0.34)
		Asiatic	5	2151	1	1	0.00019	-0.816 (0.30)
	Col-like	Worldwide	55	2099	40	16	0.00251	-1.335 (0.06)
		Asiatic	14	2105	15	7	0.00192	-0.591 (0.30)

(a) Number of sequences – (b) Length of the sequenced region (in bp) – (c) Number of segregating sites in the region (number of mutations if different from S) – (d) Number of nucleotide differences per site between two sequences – (e) Tajima's D statistics value (p-value)

skewed toward negative values [14]. Besides, this pattern is likely to be more prevalent in 'West Asia' –where the Sha allele is confined– because of the recent and rapid colonization of this region [19] (figure 25). It would have been very interesting to compare the Tajima's D values we obtained with the ones observed genome-wide but the set of accessions we used (enriched in Asiatic accessions compared to most of the sets analysed before) was very different from the Nordborg's one for which genome-wide Tajima's distribution was available. For the same reason, simulating Tajima's D using coalescent theory and taking into account previously-defined *A. thaliana* demographic models [16] was not relevant. In collaboration with C. Toomajian we looked at the haplotype sharing around MOT1. But the few Sha-like accessions (and more generally 'West Asia' accessions) that were present in the set of 408 accessions used at that time for population genetics analyses did not allow us to conclude on haplotype sharing. Besides, the genome-wide 250k SNPs set used for this analysis does not capture the SNP responsible for D104Y mutation, making the definition of the Sha-like core haplotype difficult.

It is now possible, using the 1001 genome data, to see that Sha-like accessions are associated with a strong decrease of diversity around *MOT1* compared to Col-like and Ler-like accessions in a set of accessions from 'West Asia' . This low level of polymorphism could result from a strong selective sweep (figure 26). However, the low diversity region covers about 45 genes that could also be under strong positive selection. In that case, D104Y mutation, even if not

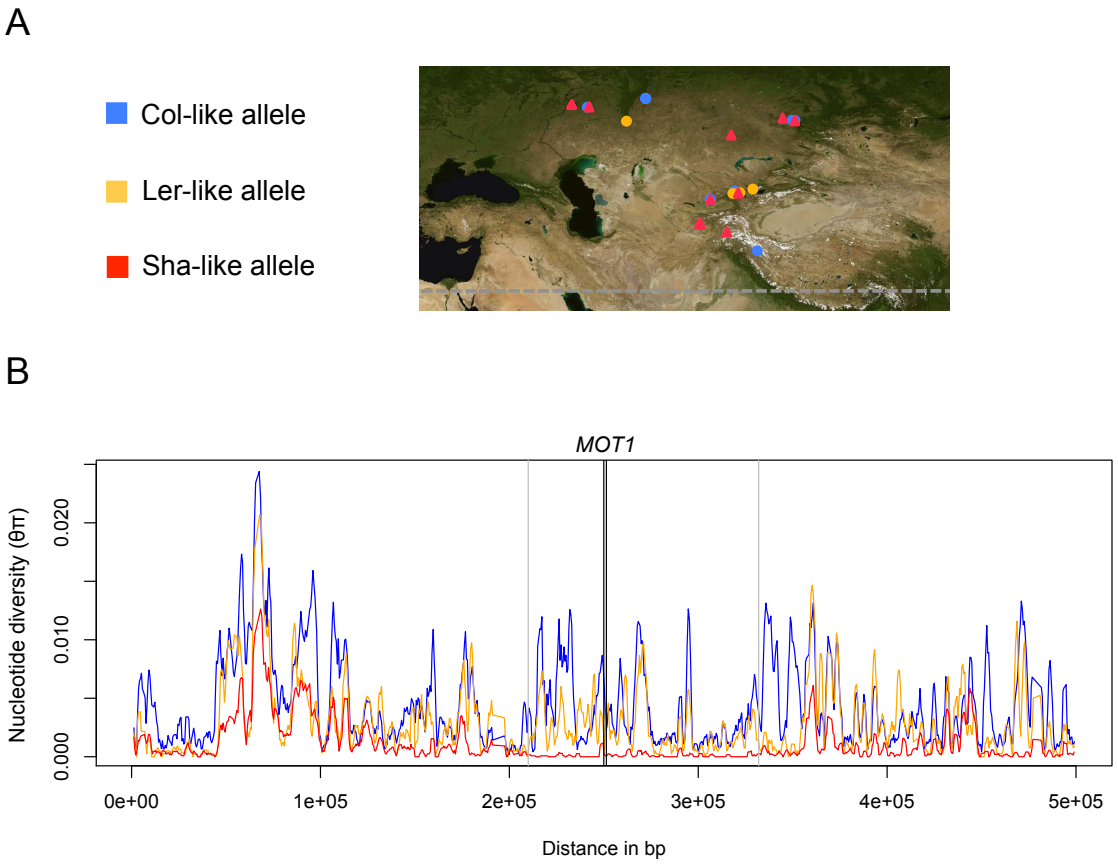


Fig. 26. Genetic diversity around *MOT1* in *A. thaliana* Col-like, Ler-like and Sha-like accessions sets in a restricted region of West Asia. A. The accessions from a restricted region of West Asia and available on 1001 genome website have been localised on a worldwide map. This set comprise 9 Col-like accessions (blue), 4 Ler-like acessions (orange) and 12 Sha-like (red) accessions. B. Sliding-window analysis of nucleotide diversity (θ_π) within Col-like, Ler-like and Sha-like sets around *MOT1* gene. Window size is 2000 bp, and step size is 500. The position of *MOT1* (black vertical line) and the Sha-like specific low diversity region (delimited by grey vertical lines) are indicated.

adaptive, could have been hitchhiked with another advantageous variant. Conversely, gene surfing could be responsible for both the pattern of polymorphism observed in *MOT1[Sha]* and the increase in frequency of this allele in 'West Asia' compared to other regions [249]. It is a strong possibility, as we know that 'West Asia' was only recently colonized. Although the exact migrant pool is not known, the presence of the Sha allele in Sweden suggests that Sha-allele did not appear during the colonisation process but was more likely at relatively high frequency among 'West Asia' migrant pool. More analyses would be necessary to properly distinguish the effect of demography and selection in the increase in frequency of the Sha-allele in 'West Asia'.

The striking occurrence of *MOT1[Sha]* on soils rich in Mo (figure 27) and the observation of Mo toxicity (Publication – Fig. 4, Fig. S7) were key elements to propose that this hypofunc-

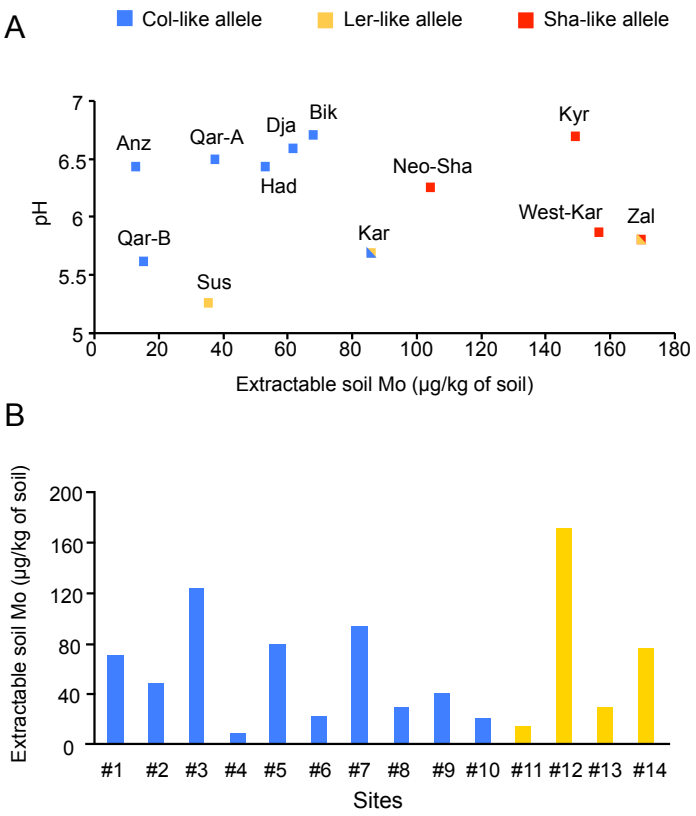


Fig. 27. Soil extractable Mo from the precise original collection site is represented for independent individuals carrying *MOT1[Col]* (blue) *MOT1[Ler]* (yellow) or *MOT1[Sha]* (red) alleles. Individuals originates from 'West-Asia'(A) or have been collected in 14 different sites by Kirsten Bomblies from around Tossa del Mar in Spain (B). No Sha-like accessions have been isolated in Spain as predicted from the worldwide distribution of Sha allele.

tional allele could be adaptive. However, as mentioned in the introduction, special care must be taken when performing correlations among different experimental and environmental data. First, because of the few edaphic data available it is important to remind that the correlation between soil Mo availability and Sha alleles could result from sampling bias (we could rely only on our own collections). More soil data, particularly for Sha-like accessions, would be necessary to reinforce this correlation.

Besides, it is difficult to know whether the level of Mo observed in 'West Asia' natural environments correspond to high levels in other worldwide locations or not. The method we used to extract elements from soils is supposed to be a measure of elemental availability but do not correspond to the total amount of elements which is often reported in soil composition databases. In a region from Spain, similar levels of Mo were found (Kirsten Bomblies, personal communication) with apparently indifferently Ler-like and Col-like accessions growing at those sites (figure 27). Consequently, we can wonder if there is a functional difference between Ler-like and Sha-like alleles (see next section), or if the conditions where Sha-like accessions are growing are really toxic regarding Mo.

Overall, Mo toxicity has been poorly documented in plants compared to animals [266]. From our experiments, high levels of Mo strongly affect root growth whereas shoot growth is not strongly affected (figure 28). Root Mo toxicity somehow reflects on fitness, as plants growing on conditions rich in Mo produce much less seeds than when they grow on standard conditions (Publication – Fig. S7). Because the Mo concentrations we used for our Mo-toxicity tests (from 1.5 g/L in watering solution) might be several orders of magnitude higher than what can be observed in the wild (on average 2.3mg/kg, [266]), it remains to be shown whether Mo toxicity is relevant in the wild and whether the levels of Mo we observed where Sha-like accessions were growing are toxic or not for the plant. However this may be difficult to test as other environmental variables could be important. For example, high levels of Mo through their effect on root growth could be particularly problematic under drought stress but may not be under well watered conditions. Finally, we did not observe a strong and significant interaction between MOT1 functionality and Mo toxicity on fitness nor on root growth in our experimental conditions. Does it mean that *MOT1[Sha]* is not adaptive regarding Mo toxicity?

One possibility is that unobserved and unexpected environmental variables combined or not to Mo toxicity, interfere and/or explain the occurrence of *MOT1[Sha]* allele on soils rich in Mo. In 'West Asia' soils sample, we observed that selenium (Se) is the only mineral which amount is correlated with the amount of Mo (figure 29). Se is known to be toxic in plants at high concentration [278]. Besides because S, Se and Mo are chemically similar, they may share common transporters. For example, SULTR1;2 can transport both S and Se [279, 280, 281]. Does MOT1 transport Se and could Se toxicity play a role in the distribution of the Sha-allele? Although we did not investigate this possibility in detail, several lines of evidence suggest that

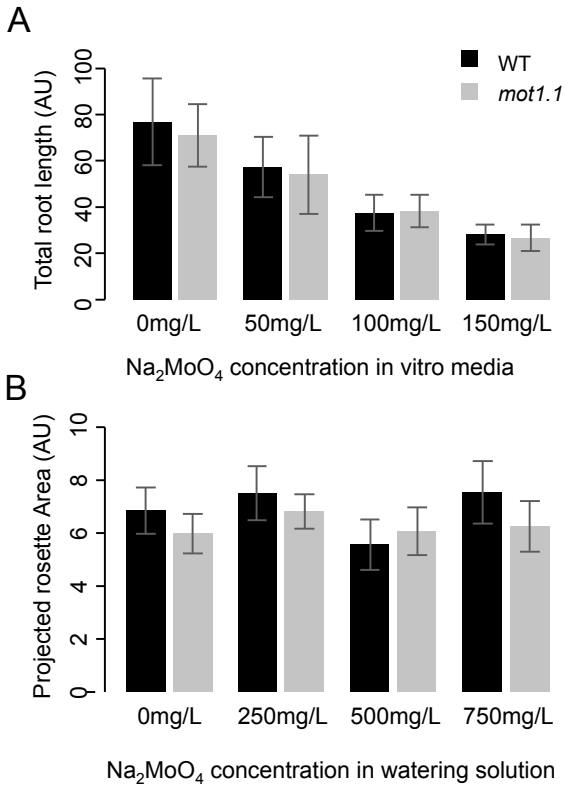


Fig. 28. Molybdate toxicity effect on root and shoot growth. A. Total root length of 11 DAS seedlings grown *in vitro* on different media supplemented with different amounts of Na_2MoO_4. B. Projected rosette area of 21DAS plants grown in soil complemented with different amount of Na_2MoO_4.AU: arbitrary unit. Error bars represent standard deviation (A) or unbiased standard deviation (B).

it is not the case. First, the level of Se is similar in tissues of Col-0, *mot1.1* mutant and Shahdara accession (ionomics database). Then, no QTL for Se tolerance has been detected around the *MOT1* region in RILs sets issued from the cross between Ler (defective MOT1) and Col-4 (functional MOT1) accessions [282]. Overall, it is still possible that a combination of environmental variables explains the occurrence of the *MOT1[Sha]* allele on soils rich in Mo in 'West-Asia'.

To conclude, these results highlight the difficulty in formally testing adaptive hypotheses in genetic backgrounds and environmental conditions that are not exactly what exists in the wild. Reciprocal transplant experiments over several generations would probably be necessary to properly test the adaptive potential of *MOT1[Sha]* allele in the wild.

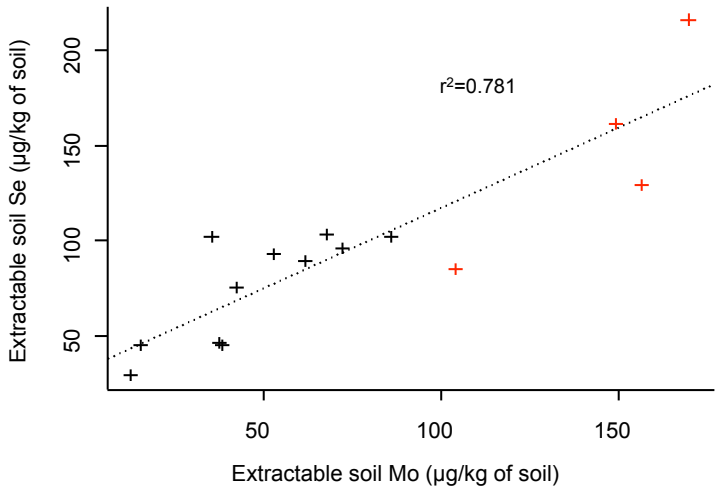


Fig. 29. Soil extractable Mo and Se from the precise original collection site is represented for independent individuals originating from 'West Asia'and carrying *MOT1[Sha]* (red) allele or not (black).

7.3 What are the differences between MOT1[Sha] and MOT1[Ler] ?

One interesting question that has not been addressed in the PLoS Genetics paper is the difference between the Ler and Sha alleles. We have shown that these two alleles were hypofunctional resulting in growth defects under soils poor in Mo. However, from an evolutionary point of view they look quite different. Indeed, the Ler allele is detected worldwide but at a relatively low frequency (figure 25). Besides, its pattern of polymorphism does not reveal any particular signature of selection suggesting that its evolution is mainly driven by genetic drift (table 4, table 5). Considering the hypofunctionality of this allele, it is likely to be deleterious on soils with low Mo availability but neutral in other environments, which could explain its worldwide maintenance. Consistent with the drift hypothesis, the occurrence of the Ler allele is not correlated with Mo availability in two independent worldwide regions (Spain and 'West Asia') (figure 27).

On behalf of the hypothesis that the Sha allele confers a selective advantage on soils rich in Mo, we can wonder why a similar adaptive pattern is not observed for the Ler allele. One possibility is that the two alleles are not functionally equivalent. Indeed, the deletion in the *MOT1[Ler]* promoter results in a reduced level of expression of *MOT1* but the functional protein is probably still present although at a lower level. Sha allele affects a given amino acids leading to a hypofunctional MOT1 transporter; all proteins are affected. These differences could result in different sensitivity thresholds, the Sha allele being more sensitive to Mo depletion than the Ler allele. This trend has been observed in some of our peatmoss experiments in which the Sha-like accessions appeared more affected on peatmoss than the Ler-like accessions

Tab. 5. Genetic diversity and population based divergence tests of selection at MOT1 in Sha-like, Van-like and Col-like subgroups.

			N (a)	Length (b)	S (c)	π (d)	Fay & Wu H (e)	MK test (f)	HKA (g)
MOT1	**Sha-like**	Worldwide	28	1428	2	0.0001	-1.79 (0.01*) (h)	1.45 (0.40)	5.948 (0.20)
		Asiatic	25	1428	1	0.00006	-1.84 (0.001***) (h)	1.35 (0.50)	8.539 (0.07)
	Ler-like	Worldwide	19	1428	4	0.00061	-1.15 (h)	1.13 (0.78)	6.012 (0.20)
		Asiatic	5	1428	2	0.00056	0.60 (h)	1.22 (0.65)	5.188 (0.27)
	Col-like	Worldwide	55	1428	43	0.00245	-22.72 (0.004***)	3.14 (0.0063**)	11.159 (0.025*)
		Asiatic	14	1428	17	0.00203	-9.278 (0.01**)	1.88 (0.14)	7.310 (0.12)

(a) Number of sequences – (b) Length of the sequenced region common between *A. thaliana* and *A. halleri* (in bp) – (c) Number of segregating sites in the region (number of mutations if different from S) – (d) Mean number of nucleotide differences per site between two sequences – (e) Fay & Wu H statisic (p-value). The tests on COI were not significant – (f) McDonald-Kreitman test using substitutions in coding regions. Neutral index (p-value of Fisher statistics). The tests on COI and PI were not significant. – (g) Multilocus HKA test was performed using MOT1, PI and COI loci. Sum of deviation (probability from Chi square distribution). The tests using COI and PI only were not significant. – (h) The low number of segregating sites in that group may bias the results.

(figure 30). However, this trend was difficult to quantify because phenotypes were more or less stable depending on greenhouse temperature and accessions' genetic background. Overall, the *mot1.1* mutant often showed a stronger phenotype compared to Ler- and Sha-like accessions confirming at least that hypofunctional alleles (expression or protein) are phenotypically not equivalent to a KO allele. If the Sha allele is more sensitive to Mo depletion than the Ler allele, it could also be expected to be more protective against high levels of Mo. Thus, it would not be surprising that the Sha-like and Ler-like alleles that are not functionally equivalent took different evolutionary routes.

7.4 Further investigations

In the last subsection, I discussed a lot about the hypofunctional alleles. However the *MOT1[Col]* alleles are probably also quite interesting to further investigate. Indeed, at the worldwide scale, McDonald Kreitman and HKA tests suggest that this set of accessions might still be diversifying at MOT1 (table 4, table 5). Thus, further advantageous or deleterious alleles under particular conditions might be segregating in *A. thaliana* natural populations. Kian phenotyped many different Col-like accessions on peatmoss and we did not observe any particular phenotypic effect differentiating those accessions. However, the phenotyping on peatmoss were probably too strong and not enough quantitative to detect possible subtle functional differences in Col-like alleles. Measuring Mo content in different accessions or populations, in different plant organs and under different environments could reveal new functionally important *MOT1* alleles. New loci involved in Mo homeostasis natural variation could also be identified. Indeed, although MOT1 has been shown to be the major regulator of plant Mo content in *A.*

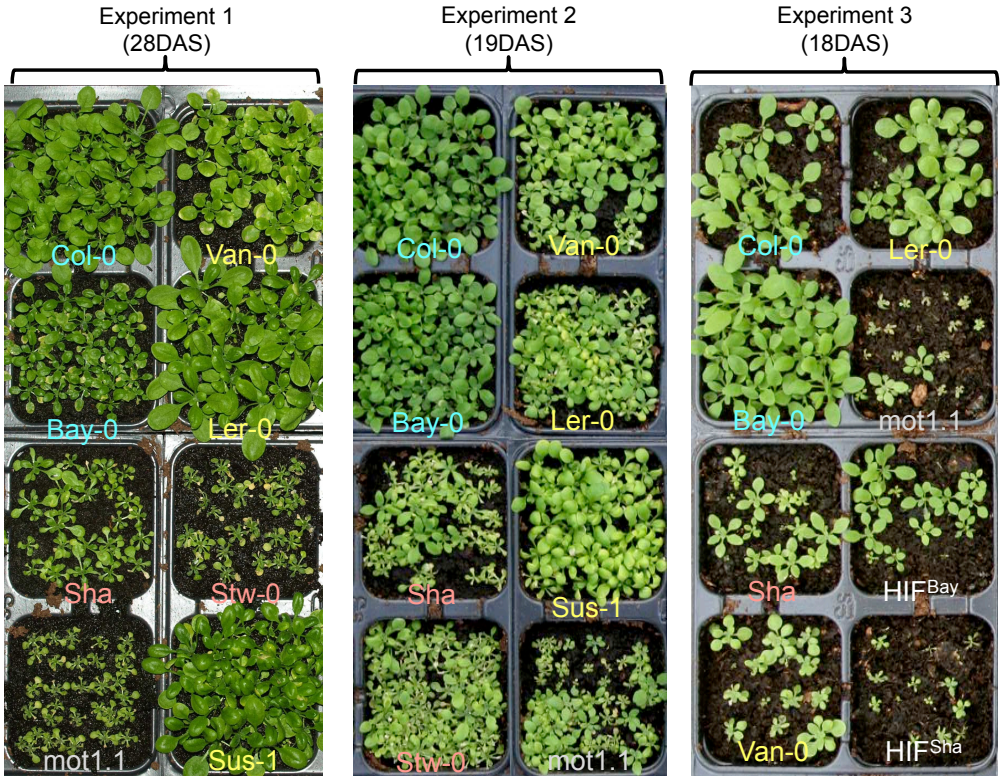


Fig. 30. Peatmoss phenotypes of different accessions in three independent experiments. Note that, despite variation between the different biological replicates, mot1.1 often looks the most affected on peatmoss, followed by the Sha-like (red), Ler-like (yellow) and finally Col-like (blue) accessions. The age of the plant is indicated for each experiment.

thaliana accessions and other species, other genes such as MOT2 could also be important [283]. Besides, Buesher and colleagues identified some other potentially interesting QTLs in Bay x Sha and Cvi x Ler RILs populations [166].

Part III

A TANDEM OF RECEPTOR-LIKE KINASES IS RESPONSIBLE FOR NATURAL VARIATION IN SHOOT GROWTH RESPONSE TO MANNITOL TREATMENT IN A. THALIANA

8. PROJECT BACKGROUND AND PERSONAL CONTRIBUTION

Since its start, the VAST group is particularly interested in decoding the genetic bases of abiotic stress tolerance in *A. thaliana* natural populations using quantitative genetics approaches. One of the issues of such strategy is the development of high-throughput phenotyping displays (see 1.3.2). In the lab, *in vitro* protocols were developed to study the effect of osmotic constraints on leaf growth at early developmental stages. One of the osmotic agents chosen at that time to perform QTL mapping was mannitol. This polyol –known to be a 'compatible solute' in the way that it should not perturb cellular macromolecules and so could be up- or down-modulated without interfering with cellular function– has been detected in more than 110 species of vascular plants. Some of them, such as celery (*Apium graveolens*), common plantain (*Plantago major*), or sour cherry (*Prunus cerasus*) accumulates mannitol to resist salt and osmotic stress [284]. However, mannitol has not been detected in *A. thaliana* and is not known to play any obvious role in this species which is why it is expected to be a relatively good osmotic agent [284]. Besides, a tremendous variation for shoot growth has been observed under mannitol osmotic stress among several accessions and RILs compared to other osmotic media such as PEG. The phenotyping of several RILs populations under control and mannitol 60mM (Man60) supplemented media was performed and several QTLs were isolated (figure 31).

One of the strongest mannitol stress-specific QTL, named *EGM* (for Enhanced shoot Growth under Mannitol stress) and isolated in Cvi-0 x Col-0 RILs population on the top of chromosome 1, was validated by Claire Le Metté using HIF and fine-mapped by Kian Poormohammad to a 10 kb region. When I arrived in the lab, Kian, using advanced recombined HIFs –only segregating for the 10kb interval– had confirmed the presence of *EGM* in this region. Besides, the analysis of T-DNA mutants in the three genes present at that locus suggested that *At1g11300* was the causative gene for EGM and that the Cvi allele of this gene was defective causing enhanced shoot growth on Man60. The goals of my PhD were to confirm the QTG, to identify the polymorphism(s) responsible for EGM and to find a link between the function of the candidate gene and the segregation of the QTL on Man60.

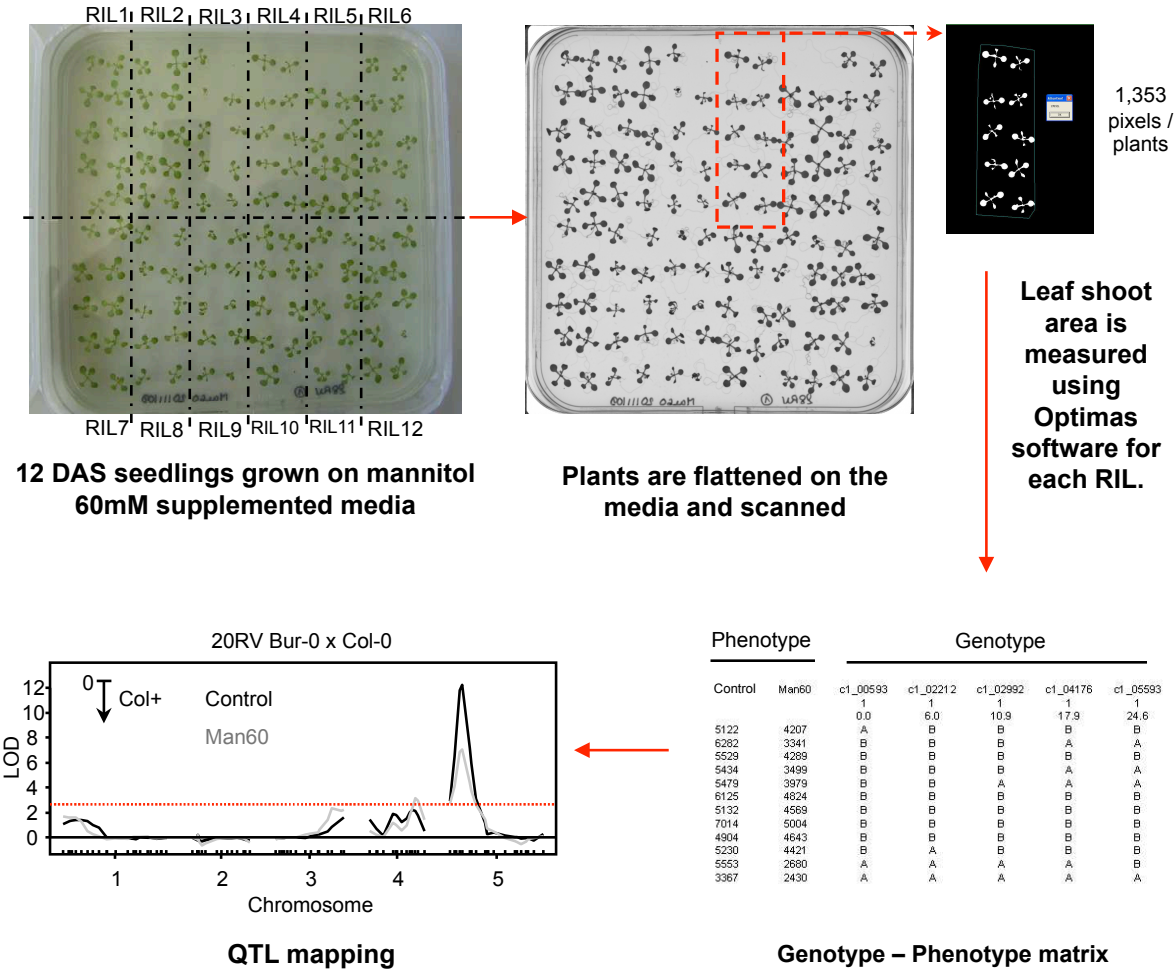


Fig. 31. *in vitro* phenotyping for QTL mapping.

9. PUBLICATION

A TANDEM OF RECEPTOR-LIKE KINASES IS RESPONSIBLE FOR NATURAL VARIATION IN SHOOT GROWTH RESPONSE TO MANNITOL TREATMENT IN *A. THALIANA*.

The 'EGM' paper will be submitted to PNAS. My contributions to this paper are highlighted in figures 1, S2, S3, 2, S4, 3, 4, S5, S6, S7A-B, TableS1, Table S2.

A tandem of receptor-like kinases is responsible for natural variation in shoot growth response to mannitol treatment in A. thaliana

C. Trontin,[1] S. Poormohammad Kiani,[1] J. Corwin,[2] K. Hematy,[1] D. Kliebenstein,[2] and O. Loudet[1]

[1] *INRA, UMR 1318, Institut Jean-Pierre Bourgin, F-78000 Versailles, France*

[2] *Univ. of California-Davis, Dept of Plant Sciences, One Shields Ave, Davis, CA-95616, USA*

Plant growth may vary considerably between and within species and this variation -as one component of fitness- is likely to reveal interesting adaptive features. However as growth is a complex trait that integrates many internal and external signals, understanding the molecular origin of this variation remains a challenging issue. In this study, natural variation for shoot growth under mannitol-induced stress was analysed by standard quantitative trait locus mapping methods in a RIL population issued from the cross between Col-0 and Cvi-0 *Arabidopsis thaliana* accessions. Cloning of a major QTL specific to mannitol condition lead to the identification of *EGM1* and *EGM2*, a pair of tandem-duplicated genes encoding receptor-like kinases potentially involved in the signalling of mannitol stress response. Using different genetic approaches, we identified two non-synonymous mutations in *EGM2[Cvi]* allele shared by at least ten accessions from different origin, likely responsible for a specific tolerance to mannitol. Indeed, we showed that the enhanced shoot growth phenotype contributed by the Cvi-allele is not linked to the osmotic stress properties of the media but more likely to the mannitol itself. This result raised the question of the function of such a gene in *A. thaliana*, a species which does not synthesize mannitol. Our findings suggest that the RLK encoded by those genes could be activated by the mannitol produced by pathogens such as fungi and contribute to plant defense response.

Introduction

Despite being physiologically disputable [1], mannitol-treated plants are still widely used as a way to induce responses in vitro and mimick osmotic constraints, including as a proxy for drought stress. Its application has helped to transcriptionally reveal pathways that are at least partially relevant to abiotic stress responses in general, but also to biotic interactions [2]. Whether this highlights cross-talks between stresses or reveals specific features of the mannitol treatment remains unclear [3]. Mannitol, the chemically reduced form of mannose, is a compatible solute that is accumulated by several plant species as a carbon storage and translocation form, and in response to abiotic stresses [4]. Despite harbouring some of the mannitol transport and enzymatic components, *Arabidopsis thaliana* is not known to accumulate mannitol and the polyol is not recognized to play any obvious role in this species [4–6]. However, it is likely that plants are exposed to external sources of mannitol in nature, as some pathogens infesting plants have been described to storing C in the form of mannitol and releasing it in planta; there is debate as to the role of this release in plant pathogen interactions [7].

Using natural variation in quantitative genetics approaches represents a no-a priori strategy to reveal new genes or alleles involved in the variation of complex traits [8], but QTL mapping outcome is of course directly conditioned by the variation segregating in the genetic material used, especially in a specific cross. Sometimes, the main sources of variation uncovered can be only very indirectly related to the initially-intended screen ([9, 10]. In this study, we took advantage of the existing RIL set between the genetically distant accessions Cvi-0 and Col-0 to map and clone a major effector of plant response to mannitol. This turns out to correspond to variation in at least one of two putative receptor-like kinases duplicated in tandem, that are directly involved in plant specific response to mannitol itself (not its osmotic effect) and in turning on several pathogen-defense components. Indeed, we describe a pathogen-tolerance phenotype associated with variation in this step.

Results

Mapping of EGM QTL.

Natural variation for seedling growth under control and mannitol-supplemented conditions has been investigated in a core set of 164 RILs issued from the cross Cvi-0 x Col-0. Among other smaller-effect QTLs, one major peak mapped to the north of chromosome 1 specifically under mannitol treatment and explained about 59% of the total phenotypic variance with a positive allelic contribution from the Cvi allele (Fig. S1A). A 2D scan of the genome revealed no interaction with any other locus. The segregation of this QTL named EGM (for Enhanced shoot Growth under Mannitol stress) was confirmed using a heterogeneous inbred family (HIF170). Direct phenotyping of the progeny of the heterozygous RIL170 (progeny testing), as well as phenotyping of the fixed homozygous lines HIF170[Col] and HIF170[Cvi], confirmed the specific segregation of the QTL under mannitol stress (Man60) conditions and revealed that the Col allele is almost completely dominant over the Cvi allele (Fig. S1B). To

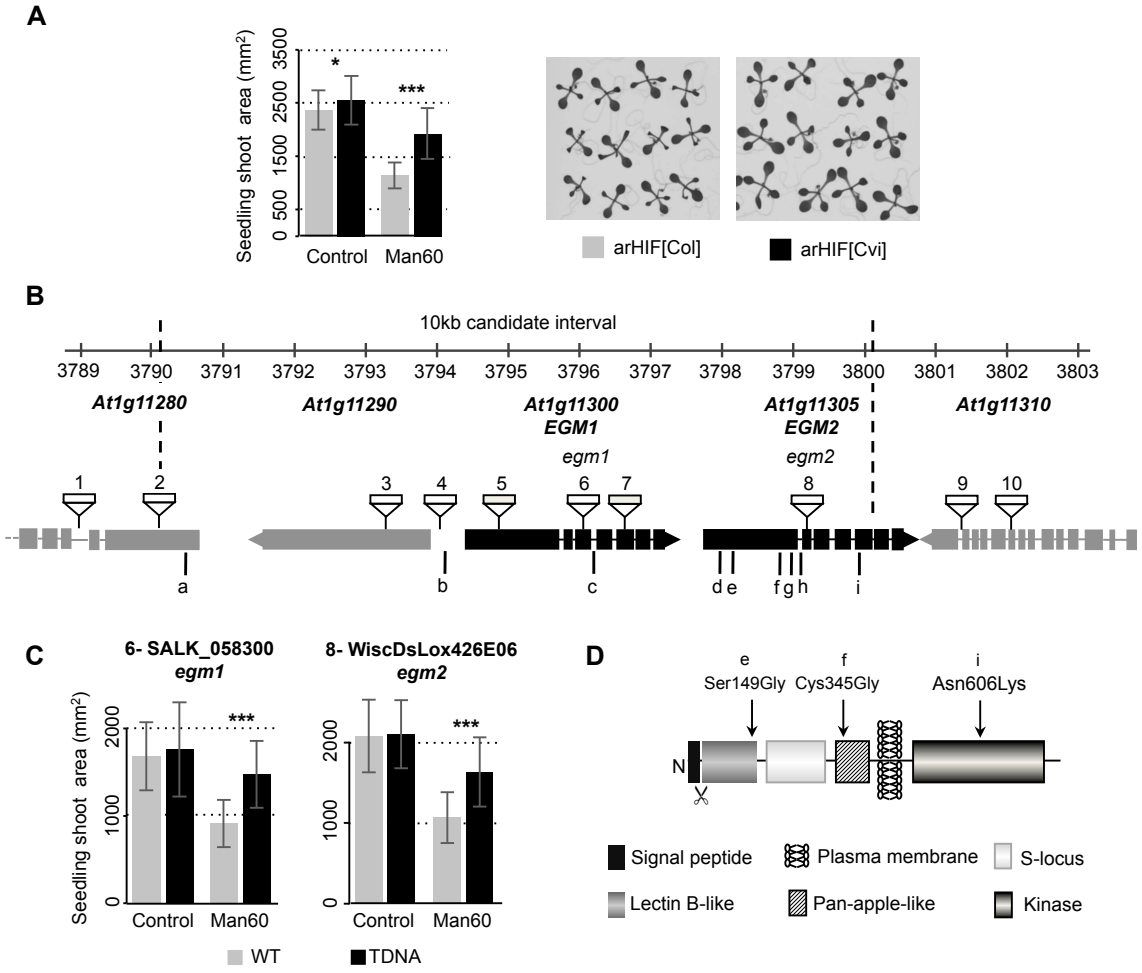


FIG. 1. **Two genes in a 10kb interval are strong candidates for EGM QTL.** A. The phenotyping of arHIF[Col] and arHIF[Cvi] 12 DAS under control and mannitol 60mM-supplemented media (Man60) confirmed the segregation of EGM QTL in a 10kb interval. Pictures illustrate the phenotype of the arHIFs on Man60. B. The 10kb candidate interval on chromosome 1 confirmed by the arHIF is delimited by dashed vertical lines. Physical positions are indicated in kb. GeneFarm-predicted gene models (arrows) are represented with exons (filled rectangles), except for *At1g11310* (TAIRv10 gene predictions). The approximate insertion sites of T-DNA in line phenotyped within the region are indicated. 1: SALK_206891; 2: WiscDsLoxHs015_10B; 3: SALK_122320; 4: SALK_008433; 5: SAIL_150_H02; 6: SALK_058300; 7: SALK_044069; 8: WiscDsLox426E06; 9: SALK_050191; 10: SALK_07985. SNPs observed between Col and Cvi alleles are indicated by letters (a-i). For the nature of the mutation see Figure 1D and 2. C. Phenotyping of *egm1* and *egm2* T-DNA mutants on control and Man60 media. D. Model of the protein structure of EGM1 and EGM2 RLKs. The peptide signals (amino-acid 1 to 27) and transmembrane region (amino-acid 439 to 461) were predicted using SignalP 3.0 server and TMHMM server v.2 respectively. The non-synonymous mutations identified in *EGM2* between the Col and Cvi allele are indicated. Error bars represent standard deviation obtained from the phenotyping of at least 30 plants and a second biological replicate gave similar results. Stars represent the significant genotypic effects obtained for each media with a Student t-test (* $0.01 < p < 0.05$; *** $p < 0.001$).

identify the quantitative trait gene (QTG) underlying EGM, fine-mapping was performed using a series of recombinants (rHIF) issued from HIF170 and ultimately reduced the candidate interval to 10kb (Fig. S1C). The analysis of advanced recombined HIF obtained from the most informative recombinants and segregating solely for the 10kb interval (Fig. S1D; Fig. 1A), confirmed that the 3,790 - 3,800kb interval of chromosome 1 is sufficient to cause the EGM-associated phenotype.

According to TAIR10, the 10kb candidate region contained three predicted open reading frames: *At1g11280* encoding a putative receptor-like kinase (RLK), *At1g11290* coding for a PRR endonuclease and *At1g11300* which was annotated as a double RLK containing the same structural unit duplicated in tandem.

Such a double structure would be quite uncommon for a kinase [11] and were indeed predicted as two different ORFs by the Genefarm algorithm (Fig. 1B). To clarify the number of ORFs encoded in the region encompassing *At1g11300*, we performed RT-PCR on the arHIF's total RNAs. It was possible to amplify each structural unit independently in both arHIFs but the amplification of a fragment encompassing both of them was not possible suggesting that the two structural units were transcribed independently. Those results were confirmed by 3'-RACE which allowed the detection of polyA tails at the end of the first and second structural unit. To sum up, four ORFs were present in the candidate interval: *At1g11280*, *At1g11290*, *At1g11300* and *At1g11305* (Fig. 1B).

Two polymorphisms likely responsible for At1g11305 hypo-functionality contribute to EGM QTL.

To identify *EGM* causative gene, several T-DNA insertion mutants in 4 positional-candidate genes were analysed (Fig. 1B). Three T-DNA insertion lines in the coding sequence of *At1g11300* or *At1g11305* and one insertion line in the promoter of *At1g11300* (affecting transcript accumulation, Fig. S2C), were associated with a coherent phenotype under Man60 conditions, i.e. lines homozygous for the insertion grew better on Man60 than the homozygous WT (wild-type) lines while showing either a significant genotype x media interaction (ANOVA ($p < 0.05$), SALK_008433, SALK_058300, WISC_DsLox426E06, SAIL_150_H02) or a much stronger effect under Man60 conditions (SALK_044069) (Fig. 1C, Fig. S2A). In addition, we generated a non-specific amiRNA line targeting the two genes and leading to a similar phenotype (Fig. S2B, Fig. S2C). This suggests that the two putative RLKs, important for plant growth response to mannitol, are good candidates for the EGM QTL. We renamed the first gene *EGM1* and the second one *EGM2* based on the phenotype of the mutants. They encode closely-related proteins (86.7% identity) from the SD1 RLK family [12]. RLKs are characterized by the presence of a signal sequence, a ligand binding extracellular domain, a transmembrane region and an intracellular carboxy-terminal kinase domain, all predicted in EGMs proteins (Fig. 1D). The SD1 subfamily is characterized by a B-lectin domain (Curculin-like domain, agglutinin motif) predicted to be involved in mannose binding [11, 13, 14], a S-locus domain and a PAN/APPLE like domain expected to be involved in protein/protein or carbohydrate/protein interactions [11, 15]. A phylogenetic analysis of the SD1 subfamily revealed that *EGM1* and *EGM2* are close paralogs that appeared by tandem gene duplication probably around the time of divergence with *A. lyrata* where only a clear *EGM1*-ortholog is identified (Fig. S3A). To check if other members of the SD1 subfamily could be involved in mannitol stress response we analysed T-DNA mutants in several SD1 members that were closely related to *EGM1*/*EGM2* and/or upregulated by mannitol treatment according to Kilian et al. 2007 [16]. Among the 8 genes we tested, we could find no obvious phenotypic indication for their implication in mannitol response (Fig. S3B, Fig. S2A).

Within the 10kb candidate region, nine polymorphisms were identified between Col-0 and Cvi-0 (a-i). Six out of nine were located in *EGM2* including three non-synonymous polymorphisms ('e'=S149G, 'f'=C345G, 'i'=N606K; Fig. 1B, Fig. 1D, Fig. 2A). To identify which of these polymorphisms was/were causal for the QTL - i.e. to identify the QTN(s)-, we tested EGM phenotypic segregation in several F2s designed to segregate for diverse combinations of Col and Cvi polymorphisms at positions 'a' to 'i', an approach we name 'specific association genetics'. This approach is not necessarily uncompromised as we cannot totally exclude that the patterns of observed phenotypic segregation could be explained by other linked polymorphisms (including those in Table S1) or influenced by other QTLs segregating specifically in each F2 population. To avoid genetic and environmental bias and possible maternal effects, we performed this approach using several independent accessions whenever possible, in independent experiments and/or on reciprocal crosses to both Col-0 and Cvi-0. First, we tested this approach on 2 accessions out of 13 presenting the same haplotype as Cvi-0 at polymorphisms 'a' to 'i' (Fig. 2A). They were redundant to Cvi-0 when crossed to Col-0 in terms of segregation of EGM, validating the approach and confirming that EGM was associated with at least one of the polymorphisms segregating in the 10kb interval. Four accessions bearing a mix of Col and Cvi polymorphisms at positions 'a' to 'i' (Bur-0, Ct-1, Shahdara and Blh-1) essentially segregated in crosses with Cvi-0 (but not Col-0), restraining the list of candidate QTNs to three (polymorphisms 'b', 'e', 'f'; Fig. 2A).

Polymorphisms 'e' and 'f' result in important amino acid changes in the Cvi version of EGM2. The first one was the mutation of a serine (conserved throughout the SD1 family) into a glycine in the lectin domain of the SD1 receptor. The second one involved a highly conserved cysteine that is predicted to be one of the six conserved cysteines of the Pan/apple-like domain involved in disulphide bonds and important for the conformation of the domain [15] (Fig. 1B, Fig. 1D). Thus, those mutations are likely to affect EGM2 functionality. To confirm this hypothesis, we complemented the arHIF[Cvi] with the Col and Cvi allele of either *EGM1* and *EGM2* under the CAM-35S promoter. Interestingly, the expression of all transgenes -but *EGM2[Cvi]*- in the arHIF[Cvi] complemented the phenotype and seedling size on Man60 was clearly negatively correlated at least with *EGM1* expression level, which overexpression was stronger than *EGM2* (Fig. 2B). EGM2[Cvi] is then at least hypo-functional probably due to the S149G and

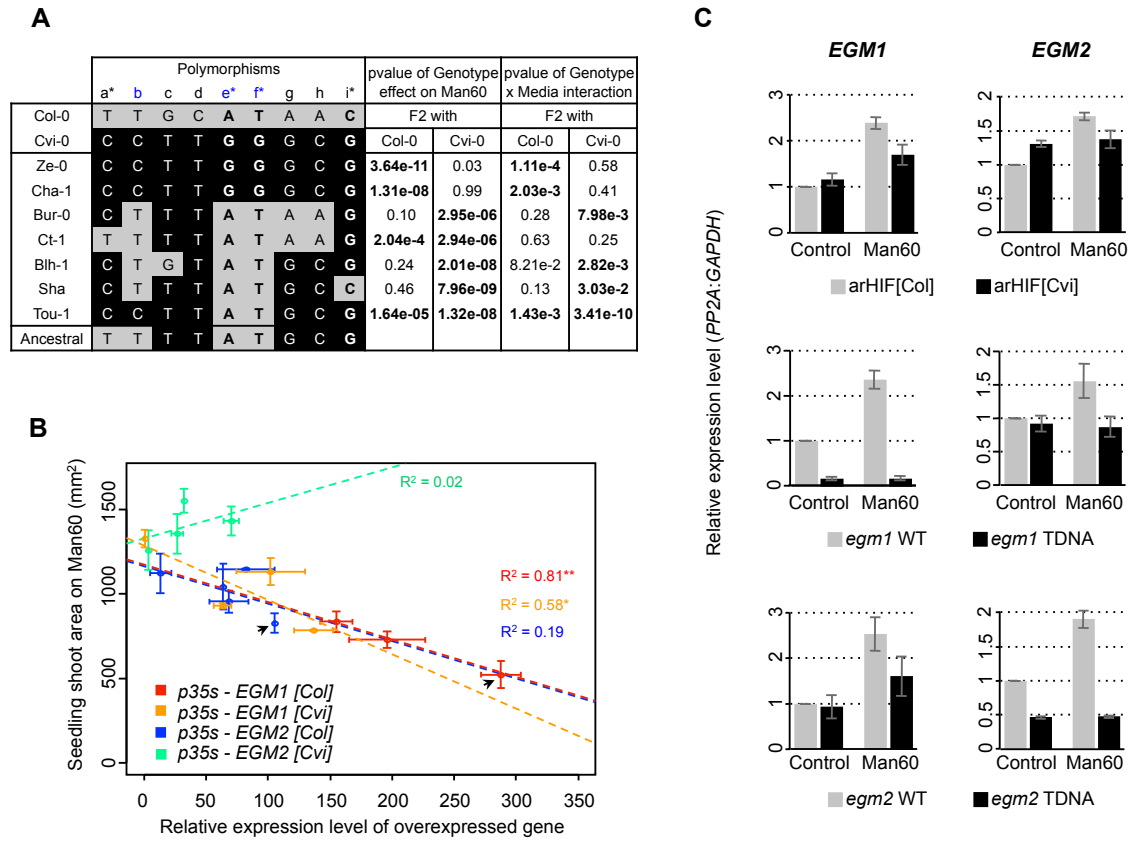


FIG. 2. **Two polymorphisms in EGM2 are likely responsible for the EGM QTL.** A. Specific association genetics identifies 3 polymorphisms ('b', 'e', 'f') as candidates for the QTL. The SNPs identified between the Col and Cvi alleles ('a' to 'i'; Figure 1B) are indicated as well as their state in 7 additional accessions. The 'ancestral' haplotype corresponds to the one retrieved from *A. lyrata EGM1* (A.ly 910885 sequence). '*' indicate non-synonymous polymorphisms (Figure 1D). The segregation of EGM in F2s issued from crosses to Col-0 or Cvi-0 has been tested through the effect (ANOVA) of the genotype at EGM on shoot area under control and Man60 conditions in at least 2 independent experiments. B. Transgenic complementation of the arHIF[Cvi] with the Col and Cvi alleles of *EGM1* and *EGM2* under the CAM-35S promoter. Arrows indicate the lines that were further analysed for mannitol stress responsive genes in Fig. S6D. Stars indicate the significance of the effect of the relative expression level of the overexpressed gene on the phenotype on Man60 (* $p < 0.05$; ** $p < 0.01$). Adjusted r-square values are also indicated. C. Relative expression levels of *EGM1* and *EGM2* in the arHIFs, *egm1* (SALK_058300) and *egm2* (WiscDsLox426E06) mutants. Expression levels were normalised with respect to the expression level of *EGM1* or *EGM2* in the arHIF[Cvi] on Man60 (B), or in the arHIF[Col] or WT under control conditions (C). Error bars represent the standard deviation observed among two biological replicates obtained from two independent experiments.

C345G mutations. Besides, the result of this experiment highlights that the level of expression of the *EGM* genes is important for shoot growth limitation in response to mannitol.

Given this result and to test if the polymorphism 'b' in the promoter of *EGM1* could also contribute to the EGM phenotype, we analysed the level of expression of both genes using specific *Taqman*® probes under control and Man60 conditions in the arHIFs. Both genes were induced under mannitol stress in the arHIF[Col]. However in the arHIF[Cvi], the induction of *EGM1* was reduced and no induction of *EGM2* could be observed (Fig. 2C). Interestingly, expression analysis in *egm1* and *egm2* mutants revealed that the induction of *EGM1* was at least partially dependent on the functionality of EGM2 and that the induction of *EGM2* was completely dependent on EGM1 functionality. In consequence, the reduction of the induction of *EGM1* observed in the arHIF[Cvi] could be explained by EGM2[Cvi] hypo-functionality. Nevertheless, we cannot totally exclude that the polymorphism 'b' in the promoter of *EGM1* could also contribute to this expression reduction as the expression level of both genes in the accession Tou-1 (Col-like at the polymorphisms 'e' and 'f' and Cvi-like at polymorphism 'b') is more similar to that of Cvi-0 than Col-0 (Fig. S4). Indeed, the Tou allele at

EGM presented an intermediate state compared to the Col and Cvi alleles in F2s (Fig. 2A). Because we only found and analysed one such accession as Tou-1, those results could also be explained by additional specific polymorphisms in Tou-1.
Overall, our results strongly suggest that at least two non-synonymous polymorphisms in *EGM2* are responsible for the QTL.

EGM1 and EGM2 act together in mannitol stress response.

Although phylogenetically close, *EGM1* or *EGM2*'s single mutants present enhanced shoot growth under mannitol treatment, suggesting that they are not fully redundant (Fig. 1C). Conversely, a p35s:*EGM1* construct was able to complement the arHIF[Cvi] that carries an hypo-functional allele of *EGM2* (Fig. 2B), which suggested that the two genes might be functionally equivalent. To check whether *EGM1* could really substitute *EGM2* even when expressed at endogenous expression levels, we compared the complementation of *egm2* mutant with *EGM1* and *EGM2* genes under the control of their endogenous promoters (about 1kb upstream ATG). Whereas *pEGM2:EGM2* construct phenotypically complemented the *egm2* mutant in the four insertion lines analysed (Fig. 3A), *pEGM1:EGM1* only complemented *egm2* in one line out of four (line A, Fig. 3A), while this construct was functional to complement *egm1* mutant. Transcript accumulation analyses revealed that the level of expression of *EGM1* in the *egm2 pEGM1:EGM1* insertion lines B, C and D was equivalent or higher than that of *EGM2* in the *egm2 pEGM2:EGM2* lines, so complementation difference between the two constructs is not just explained by a difference in their expression level, while the phenotype of insertion line A may be explained by a higher level of transcript accumulation (Fig. 3A). Those data suggest that *EGM1* can complement *egm2* only when expressed at a quite high level and that these two paralogs in Col-0 are not fully redundant.
The difference between *pEGM1:EGM1* and *pEGM2:EGM2* gene could also be explained by different expression patterns. To analyse this, the promoters of *EGM1* and *EGM2* fused to the GUS reporter were transformed into Col-0. Whereas almost no GUS staining could be detected in 15 days-old seedlings grown under control conditions, a signal was observed in the shoot meristem and young leaves of seedlings grown under Man60 condition in at least 5 out of 7 and 3 out of 9 insertion lines respectively for *pEGM1:GUS* and *pEGM2:GUS* (Fig. 3B). Some lines also revealed coloration in the petiole of cotyledons and in the hypocotyle. This pattern of expression suggested that these two genes were acting early during leaf development to induce growth inhibition under Man60 condition.
Finally, the complementation of the *egm2* mutant with the 8kb-region from the 1kb-promoter of *EGM1* toward the end of *EGM2*, lead to an increase of the level of expression of the two genes and to a stronger growth reduction on Man60, suggesting that the two RLKs act together to induce EGM response and are limitant for this pathway leading to growth reduction. The phenotyping of two independent insertion lines transformed with an amiRNA construct targeting both genes also suggested that EGM1 and EGM2 are involved in the same pathway (Fig. S2B). As a conclusion, *EGM1* and *EGM2* are not fully redundant paralogs expressed in overlapping tissues and probably act in a similar way to reduce growth under mannitol treatment.

Toward EGM function.

In this analysis, we used mannitol in our in vitro media to induce a response. However, because *At1g11305* contains putative mannose-binding and carbohydrate-binding domains, we wondered whether the variation in shoot growth results from the osmotic stress imposed by mannitol or from another action of mannitol itself.
To answer this question, we tested the segregation of EGM under different osmotic constraints such as those generated with NaCl, KCl, mannose and sorbitol-supplemented media. No significant phenotype was associated with deficient EGM alleles (Cvi or T-DNA insertion alleles) under any of these conditions (Fig. 4A; Fig. 5A). The segregation of the QTL in the arHIF background was also tested under drought stress on soil plugs at later stages of development but no significant genotypic effect or genotype x drought interaction could be observed (Fig. 4B). Interestingly, the growth difference observed between the arHIFs is established from relatively low mannitol concentrations ($\leq$ 10 mM) and is then maintained on higher mannitol concentrations while the shoot size continues to decrease probably due to the osmotic stress component (Fig. S5B). Those results suggest that EGM segregation is specific to mannitol treatment and raise the question of the function of EGM1 and EGM2 RLKs in a species which does not naturally synthesize mannitol.
To understand the link between those proteins and mannitol response, the transcriptomes of the arHIF[Col] and arHIF[Cvi] on Man60 media were compared using CATMA microarrays. 221 genes were differentially expressed between the two arHIFs, most of which (199) being upregulated in the arHIF[Col] compared to the arHIF[Cvi] (Table S2). Biological GO analysis of this set of genes revealed a significant enrichment in genes involved in response to stimulus including genes responding to abiotic and biotic stimulus (Fig. S6A). The comparison of our transcriptomic analysis with the one of Skirycz et al. who studied the response of Col-0 seedlings to mannitol 25mM treatment [2] showed a significant overlap (non-parametric randomization

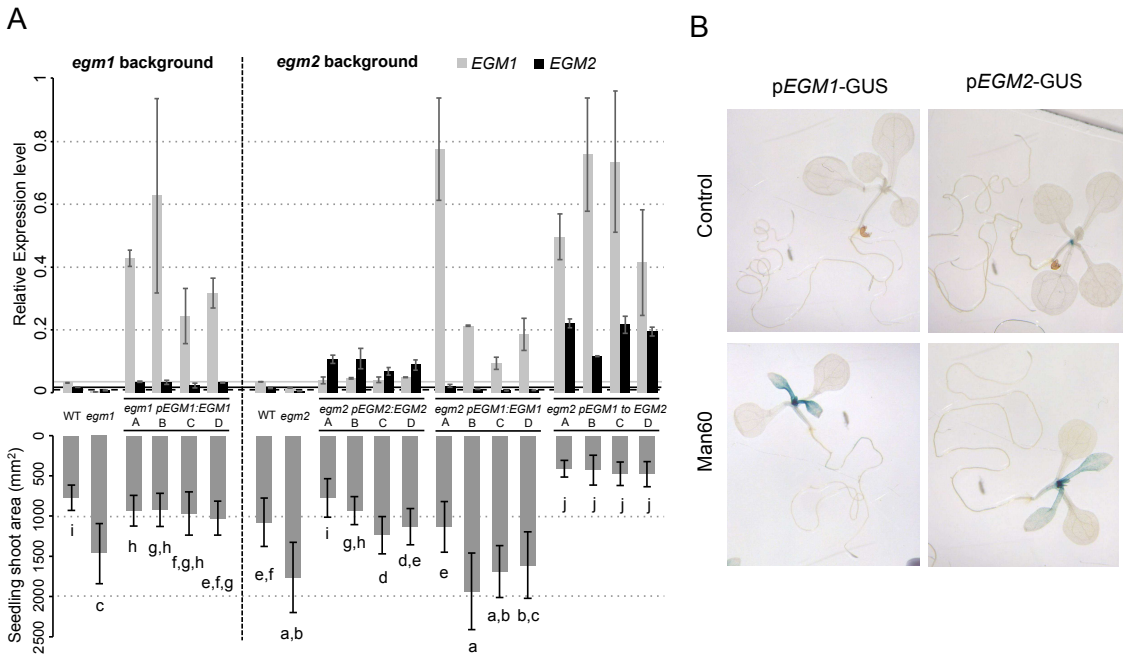


FIG. 3. **EGM1 and EGM2 are not fully redundant proteins involved in mannitol stress response.** A. Complementation of *egm1* and *egm2* mutants with *EGM1* or *EGM2* genes under the control of their endogenous promoters (1kb upstream ATG) or with the region covering the 1kb promoter of *EGM1* toward the end of *EGM2* (8kb). The shoot area and the expression of *EGM1* and *EGM2* was measured on 12 DAS seedlings grown on Man60. Different letters indicate significantly different lines based on a post-hoc Kruskal-Wallis test ($p < 0.05$). Error bars represent the standard deviation observed in two biological replicates obtained from two independent experiments (expression level) or the standard deviation obtained from the phenotyping of at least 30 plants (seedling shoot area). A second biological replicate gave similar results. B. Histochemical analysis of GUS reporter gene expression driven by *EGM1* or *EGM2* promoter (1kb upstream ATG) in 12 DAS whole seedlings grown under control or Man60 conditions.

test: $p < 0.001$): 185 of the 221 differentially-expressed genes between the arHIFs were also interrogated in Skirycz et al., among which 70 were also found to be mannitol-responsive genes (Fig. 4C). We specifically validate this observation by qPCR on 7 out of 9 tested genes that were common between the two analyses (Fig. 4D, Fig. S6B). Among the ones we confirmed, 3 genes are typical to the biotic stress response GO class and 2 were chitinases. Those genes were induced essentially specifically -or at a much higher level- on Man60 (and not on media supplemented with 80mM sorbitol, another osmoticum) in the arHIF[Col] with respect to the arHIF[Cvi]. We further validated that this induction was dependent on both EGM1 and EGM2 functionality in the mutants and amiRNA lines on 3 genes (Fig. 4D, Fig. S6C). Those results at the transcriptomic level mirored the growth phenotypes and suggested that the defective EGM genotypes (arHIF[Cvi] and *egm* mutants) behave like plants that do not perceive and respond to mannitol. Besides, in the overexpressors, the transcriptomic response was opposite than in the defective genotypes, consistent with a stronger phenotypic response in the two analysed lines (Fig. S6D). Like in Skirycz et al. [2], several genes involved in ethylene signalling were differentially expressed under Man60 between the two arHIFs (however we were not able to validate these mild differences by qPCR; Fig. S7A). To further analyse the potential role of this hormone in EGM phenotype, we crossed *egm1* and *egm2* mutants with mutants of the ethylene pathway (*ein3-1*, *ebf1-3* and *erf5*). This shows that the EGM phenotypic response at least doesn't fully rely on the ethylene pathway (Fig. S7B).

The specificity of the mannitol response with respect to other osmotic stresses at both the phenotypic and transcriptomic level combined with the enrichment in genes belonging to the biotic stress GO category in our CATMA analysis suggested that EGM might be involved in biotic stress response. We know that some pathogens including fungi use mannitol as carbon storage. Besides, during infection, it was suggested that mannitol could be secreted by pathogens to counteract plant ROS production and that plants could react against this phenomenon [17]. To test the role of EGM1 and EGM2 in plant defense against pathogens, we evaluated *egm* mutants' sensitivity to Botrytis using

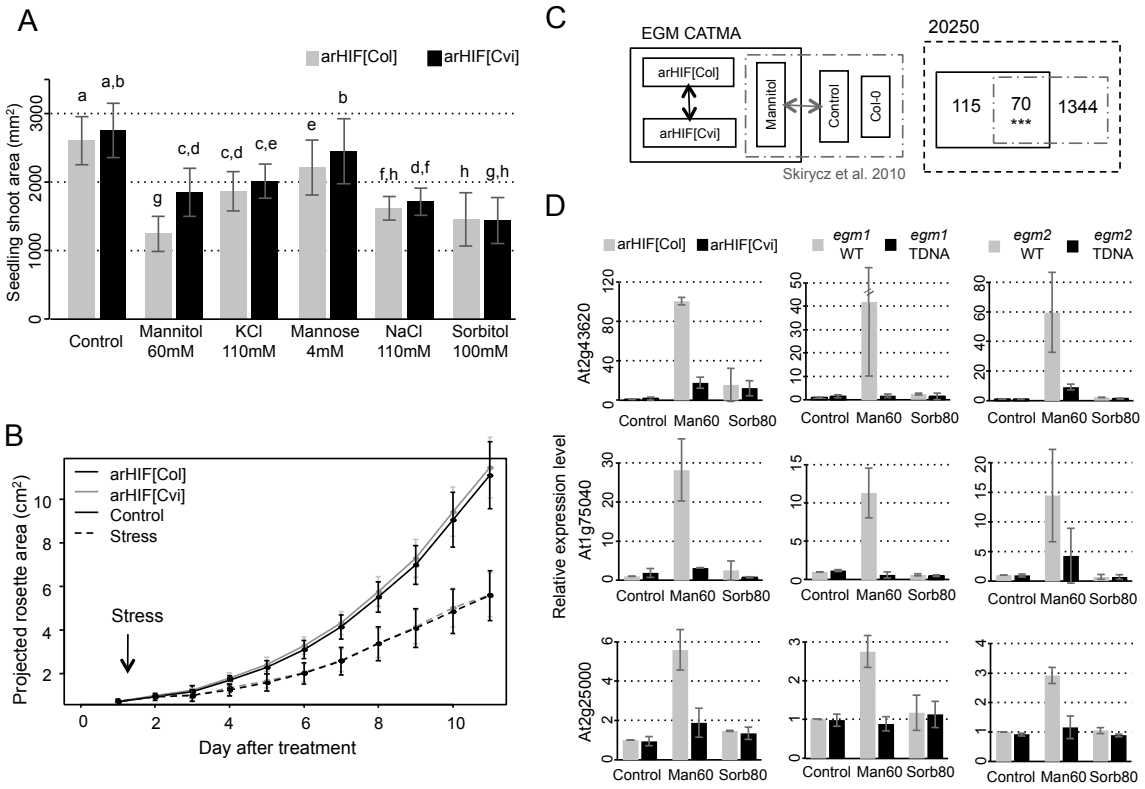


FIG. 4. **The enhanced shoot growth phenotype observed in *egm1* and *egm2* mutants on Man60 is likely not the result of an osmotic constraint.** A. Phenotyping of 12 DAS seedlings of the arHIF[Col] and arHIF[Cvi] grown under different osmotic constraints. Different letters indicate significantly different groups based on a post-hoc Tukeys HSD test ($p < 0.05$). A second biological replicate gave similar results. B. Phenotyping of the arHIF[Col] and arHIF[Cvi] on soil plugs under control (60% of plug soil water content) or drought stress (30% of plug soil water content) conditions for 10 days. At day 0, plants were already 14 days old. Error bars represent the standard deviation obtained from the phenotyping of at least 30 plants with a second biological replicate giving similar results (A-B). C. Comparison of EGM set of genes differentially expressed between the two arHIFs under Man60 with the set of genes previously identified (Skirycz et al. 2010) as differentialy expressed in expanding cells of Col-0 accession grown under control or mannitol 25mM supplemented media. The overlap between the two sets that comprises genes that were differentially expressed in the same direction in the two analyses was tested using Genesect tool available on VirtualPlant1.3 website. The 20,250 TAIRv10 gene models common between the CATMAv5 and Affymetrix ATH1 arrays were used as the background list. D. Relative transcript accumulation of three genes shared between our transcriptomic analysis and the one of Skirycz et al. (2010) in 12 DAS seedlings of the arHIFs and *egm1* / *egm2* mutants grown under control, Man60 and Sorb80 (sorbitol 80mM-supplemented media) conditions. At2g43620: chitinase, At1g75040: PR-5, At2g25000: WRK60. Error bars represent the standard deviation obtained from two biological replicates obtained from two independent experiments.

two Bc isolates (BcGrape and Bc83-2). *egm1* and *egm2* mutants were more sensitive to the two Bc isolates than Col-0, as shown by a larger perimeter of the necrotic region 72 hours post infection (Fig. 5A); similar results were found with the arHIFs despite a higher intrinsic sensitivity of this genetic background (Fig. S7C). A significant interaction between plant genotype and Bc isolates could be observed (ANOVA, $p < 0.001$), with a stronger response to the BcGrape isolate that produced significantly larger amounts of mannitol within the plant lesion (Kruskal-Wallis test, $p < 0.001$; Fig. 5B, Fig. S7D), as if EGM contribution to pathogen defense indeed depended on pathogen mannitol secretion.

Discussion

In this study, we identified two putative receptor-like-kinases involved in shoot growth repression specifically under mannitol stress and not on other osmotic stresses (Fig. 4, Fig. S5). As a consequence, our results suggest that mannitol does not only generate a generic osmotic stress but could also act as a specific signal ultimately transduced by EGM1 and EGM2. This hypothesis has already been suggested before: indeed, in several non-mannitol producing organisms transformed with either

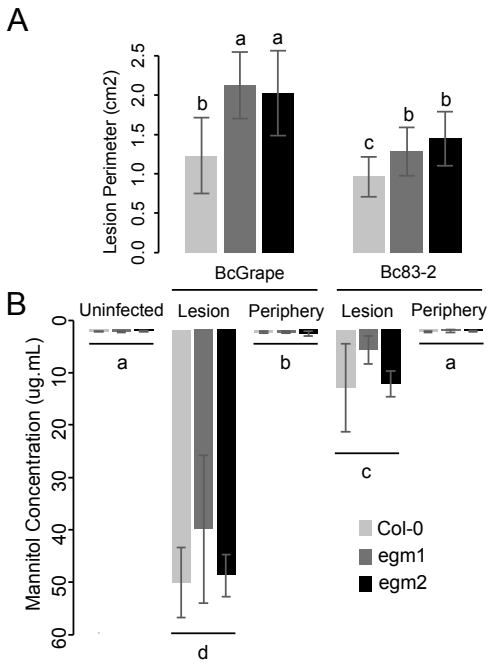


FIG. 5. **The role *EGM1* and *EGM2* in biotic stress response.** A. Mean perimeter of necrotic lesions formed by *B. cinerea* isolates BcGrape or Bc83-2 on wild-type, *egm1* and *egm2* mutants plants 72 hours post-inoculation. Different letters indicate significantly different groups based on a post-hoc Tukeys HSD test ($p < 0.05$). B. Mannitol concentrations observed within or at the periphery of necrotic lesions formed by B. cinerea isolates BcGrape or Bc83-2 on wild-type, *egm1* and *egm2* mutant plants 72 hours post-inoculation. Different letters indicate significantly different groups based on a post-hoc Kruskal-Wallis test ($p < 0.05$).

bacterial mannitol-1P-deshydrogenase (*mt1d*) or celery mannose-6-phosphate reductase (*M6PR*), enhanced resistance to salt, osmotic and/or drought stress have been observed [18–24]. Interestingly, the amount of mannitol produced in all those lines was too low to protect against stress through osmotic adjustment suggesting that mannitol could have other stress protective functions. In 2011, Chan et al. compared the transcriptomic response of *M6PR*-overexpressing *A. thaliana* lines with Col-0 in control and 100mM NaCl-stress conditions [3]. As expected from phenotypic analysis, Col-0 was transcriptionnaly more affected by salt than *M6PR*-overexpressing lines. However, surprisingly considering the absence of phenotypic effect, *M6PR* transgene altered the expression level of 2272 genes in control condition with enrichment in genes belonging to 'Response to stress', 'Response to biotic and abiotic stimulus' and 'Signal transduction' GO categories. Thus, *M6PR* expression and mannitol synthesis seemed to cause preadaptive changes facilitating response to abiotic stress. In addition, more genes involved in pathogen defense were affected by *M6PR* transgene in control condition compared to salt stress in Col-0 [3]. Skirycz and colleagues also identified several biotic stress genes that were upregulated in proliferating and expanding leaves under mannitol treatment. Besides, the mannitol transcriptomic response in those tissues showed a significant overlap with publicly available microarray analysing biotic stress response [2].

The deregulation of several pathogen defense responsive genes under mannitol stress and by *M6PR* transgene could result from crosstalks between biotic and abiotic stress responses via antagonistic and synergistic relationships between ABA and SA/JA/Et hormones and/or ROS generation and signaling [25]. Skirycz and colleagues interpreted the upregulation of those biotic stress responsive genes as the result of the activation of the ethylene signalling pathway [2]. Our transcriptomic analysis also suggested that ethylene could be important for the expression of EGM phenotype (Fig. S7, Table S2). However, the phenotyping of *egm-ein3*, *egm-ebf1* and *egm-erf5* double mutants did not confirm this assumption.

Conversely, the deregulation of several pathogen defense responsive genes under mannitol stress could have a real biological significance. In our analysis, we show that *egm1* and *egm2* mutants are more sensitive than WT to two Botrytis strains and that sensitivity correlates with the amount of mannitol produced by the different strains in the plant lesion (Fig. 5, Fig.S7). Although this correlation between mannitol and pathogenicity could be fortuitous, it suggests that mannitol produced and secreted by some pathogens during infection could be perceived more or less directly by the plant. In nature, plants likely face high concentration of mannitol during infections with some mannitol-producing pathogens (Fig. 5, Fig.S7; [26]). In fungi, among other roles, mannitol is important for carbohydrate storage as revealed by its relatively high concentration [27, 28]. Although this role may not be essential for all fungi [28] it is particularly interesting for pathogens infecting plants that are not producing or metabolizing mannitol because fungi can then safely sequester the carbon from the host in the form of mannitol [26, 29–31]. Besides, several studies show in vitro and in vivo that mannitol likely plays a role in scavenging free radicals generated during stresses [26, 32–38]. Those mannitol properties might play a central role during host/pathogen interaction because hosts are often producing ROS species to face pathogen attacks [39]. As a result, it has been hypothetised that pathogens secrete mannitol to counteract host ROS production. This hypothesis is supported by the increase in mannitol production and secretion by the pathogen (*Cladosporum fulvum* [29], [17]) and/or the upregulation of mannitol synthesis enzymes during infection or treatment with plant leaf extract (*A. Alternata* [7], *U. fabae* [26]). Besides, some fungi mutated for mannitol biosynthesis

genes resulting in strong reduction in mannitol content have been shown to be less virulent than WT (*A. Alternata* [7], *Cryptococus neoformans* [36]). The role of mannitol in host/pathogen interactions is further supported by the fact that several plant mannitol deshydrogenase (MTD) (that converts mannitol to mannose) respond to biotic stress signal. For example, celery *MTD* is upregulated in response to SA [40, 41, 43]. Besides, in tobacco, that does not produce mannitol, an endogenous MTD which activity is upregulated by fungi and inducers of plant PR proteins (INA, H2O2) has been identified [17]. Finally, under SA treatment, the celery MTD overexpressed in tobacco is secreted in the apoplasm where fungi mannitol is also partly localised. Because those transgenic tobaccos showed enhanced resistance to *A. Alternaria*, a mannitol secreting pathogen and not to *Cercospora nicotianae*, a non-mannitol secreting fungal, [42] it has been hypothetised that to protect plant's ROS-mediated defenses, plant MTD is secreted in response to SA in the apoplast where it converts the fungal produced ROS-quencher mannitol to mannose [43].

Given the potentially important roles of mannitol for some fungi during plant infection (carbohydrate and ROS sequestration) it is possible that plants developed specific mannitol-sensing pathways to induce their response. Consistent with this hypothesis, we identified two SD1 receptor-like-kinases possibly contributing to pathogen defense via the direct or indirect perception of mannitol. RLK receptors are known to be involved in a wide range of developmental processes such as hormone perception, reproduction, meristem regulation and cell/organ specification, as well as in the response to biotic and abiotic stresses [11, 44, 45]. The most famous member of the SD1 subfamily is the *SRK* gene which is the determinant of self incompatibility specificity in Brassicaceae. However, this family has also been previously identified as enriched in genes upregulated by several of the biotic stresses available in AtGenExpress data [16] and by some abiotic stresses including UV-B [12]. Besides, several close homologs of EGM RLKs in A. thaliana (*At1g11350* (*CbRLK1*) [46], *At1g65790* (*ARK1*), *At4g21380* (*ARK3*) [47] and in other species (tobacco (*Nt-Sd-RLK*[48], *NgRLK1*[49]), *rice (Pi-d2)*[50]), are induced under SA, wounding and/or bacteria-generated stresses and so are likely involved in biotic stress response. Finally, this subfamily of receptor significantly contributes to the expansion of the RLK/Pelle gene family, an expansion that could be explained as a response to fast-evolving pahogens. As a result, it is very likely that EGM1 and EGM2 are also involved in biotic stress response.

Although a direct proof of mannitol perception by the plant is still missing, the sensing of mannitol associated to the perception of other microbe associated molecular patterns could contribute to the quantitative resistance of plants against mannitol-producing pathogens.

Materials and Methods

Plant material. All accessions and most of the mapping populations used in this study are from the Versailles Arabidopsis Stock Center (INRA IJPB; http://publiclines.versailles.inra.fr/) or have been generated (Table S4). For the 'specific association genetics' approach, Cvi-like accessions were identified from a screen of about 500 accessions from VASC with two CAPS markers for the polymorphisms e and f (Table S3) and from the analysis of the 1001 Genomes sequences of about 400 accessions (http://www.1001genomes.org/, http://signal.salk.edu/atg1001/index.php). All mutant lines were ordered from NASC (Table S4). The primers used to genotype all the lines are indicated in Table S3.

Shoot growth estimations. Seeds were sterilized for 10 min in 70% EtOH with 0.1% of Triton X-100 and rinsed for 10 min in 95% EtOH. They were stratified in 0.1% agar at 4C in darkness for 3 days. Plants were sown on typical Arabidopsis media (2.5mM KH2PO4, 2mM MgSO4, 2mM CaCl2, micro-elements (70mM H3BO3, 14mM MnCl2, 0.5mM CuSO4, 0.2mM Na2MoO4, 10mM NaCl, 1mM ZnSO4, 0.01mM CoCl2), vitamines (27.7mM Myo-Inositol, 4mM Niacine, 2.4mM Pyridoxine, 1.5mM Thiamine HCl, 0.21mM Biotine, 0.5g/L Panthotenate Ca), 0.8 BCP, 0.07% MES, 0.005% Fer Citrate Ammo, 5mM KNO3, 2.5mM Ca(NO3)2, 1% sucrose and 1% phytablend) and on the same media supplemented with the indicated concentration of mannitol (2.5 to 100 mM), KCl (110mM), NaCl (110mM), sorbitol (80mM-100mM) or mannose (4mM). Shoot estimates were performed as described previously [51], 12 days after sowing. For the 'specific association genetics' approach, progeny testing of the segregation of EGM was tested in F2 populations from diverse crosses in the same conditions as above. The segregation of EGM in a given cross was estimated from the contribution of the genotype effect on seedling shoot area on Man60 only (ANOVA: Area Genotype x Experiment x Cross) and from the contribution of the genotype x media component (ANOVA: Area Genotype x Media x Experiment x Cross). For in vivo (soil-based) drought stress experiment, seeds were stratified in darkness at 4C for 3 days. Fertiss plugs (filled with a mix of peatmoss soil and vermiculite) were saturated with nutritive solution and individually weighed. After 12 days on plugs at 80% of their saturated weight, 60 homogeneous plants per genotype have been selected. At day 13 plugs were let dry at 60% of their saturated weight. From now on, water content in the soil was checked everyday by weighting each plug individually. From day 14 to 16, half of the plugs were not watered to let them dry at 30% of saturated weight. Plugs were then maintained at 60% (control condition) and 30% (mild drought stress condition) during 8 more days. From day 14 to 25, pictures of the plants were taken. Total leaf area of each plant has been estimated using

Image J. [52]

QTL mapping and fine-mapping. For QTL analysis, seedling shoot area for each of the 164 RILs of the Cvi-0xCol-0 RIL set ([53]; RIL set 8RV from VASC) was estimated from the measurement of 9 plants grown on control media and mannitol 60mM-supplemented media. MQM and 2D scan analyses were performed using the R-qtl package implemented in R (http://www.r-project.org). EGM R2 was estimated using an ANOVA at marker c1_04176. Significance threshold ($p < 0.05$) was estimated using a 1000-permutation test. EGM QTL segregation was confirmed in a Heterogeneous Inbred Family (HIF) derived from the RIL170. For fine-mapping about 6,000 descendants of the heterozygous HIF170 were screened for recombinants (rHIF). By analysing the segregation of EGM in informative rHIFs, the candidate interval for the *QTL* was reduced to 10kb. Advanced recombined HIF segregating solely for the 10kb interval were obtained by fixing rHIF #40 and #59 for the Cvi allele and by crossing these two fixed recombinants.

Vectors constructions and plant transformations. All the combinations of *EGM1* and *EGM2*, with or without stop-codon and promoters (about 1kb) (except *pEGM2:EGM2* and the 8kb region from *pEGM1*to*EGM2*) were obtained by PCR amplification using Phusion® high fidelity taq polymerase (Finnzymes) and the couples of primers containing recombination sequences indicated in Table S3. Fragments were cloned in the pDONR207™ entry vector (Invitrogen™) via BP recombination according to GATEWAY cloning procedure (Invitrogen™) and subsequently transferred into various destination vectors (Table S3) via LR recombination reaction. To obtain the pDONR207:p*EGM2:EGM2* construct, a XhoI/Eco0109I fragment (1.494kb) from pDONR207:p*EGM2:EGM2*ExtraCellulardomain was ligated to a XhoI/EcoO109I fragment (5.863kb) from pDON207:*EGM2*(+stop) using T4 DNA ligase according to manufacturer's protocol (Fermentas). To obtain the pDONR207:p*EGM1*to*EGM2* construct, an AlwNI fragment from pDONR207:p*EGM1:EGM1* (4.773kb) was ligated to an AlwNI fragment from pDONR207:p*EGM2:EGM2* (5.969kb) using T4 DNA ligase according to manufacturer's protocol (Fermentas). The artificial miRNA was designed using the WMD online tool (http://wmd.weigelworld.org/) against a 21nt sequence conserved between *EGM1* and *EGM2* but with less than 72% identity with other members of the SD1 subfamily of RLK. Amplification of the amiRNA was performed using pRS300 vector as recommended on WMD website and cloned into pTOPO® (Invitrogen). The amiRNA was then placed under the control of the constitutive CaMV 35s promoter by the cloning of the BamH1/EcoRI insert from pTOPO:amiRNA into pHannibal [54] using T4 DNA ligase (Fermentas). Finally, the Not1 digested insert from pHannibal:amiRNA was cloned into the pGREENII0000 plant vector [55].
All constructs in plant destination vector have been transformed in electrocompetent C58C1 *Agrobacterium tumefaciens* which were then used for agroinfiltration of the genotypes of interest (Table S3).

Quantitative PCR. Total RNA of 11 days-old seedlings was extracted using RNeasy Plant mini kit (Qiagen) and treated with RNases-free DNAse I (Fermentas). DNA contamination was verified by PCR on at least 37.5ng ARN and first-strand cDNA was synthesized from 750ng RNA using RevertAid™ H Minus Reverse transcriptase (Fermentas) with oligo(dT)18 in 20microL reactions. The expression level of *EGM1* and *EGM2* was analysed using TaqMan® Gene Expression Assays (*EGM1*: At02283577_g1 – *EGM2*: At02283571_g1 – *PP2A*: At02284835_g1 – *GAPDH*: At02284919_g1) and TaqMan® Gene Expression Master Mix (Applied Biosystems). Probe specificity was verified using specific DNA matrices. Other analyses were made using MESA GREEN qPCR MasterMix (Eurogentec) and the couple of primers indicated in SuppTable 1. For the two different techniques, 4 and 5 microL of 5 times diluted cDNA were used and reactions were performed as manufacturers protocol on Biorad CFX96™ Real-Time PCR detection machine. For each sample we analyse at least two technical x two biological replicates. The expression of each target was normalised with PP2A and GAPDH endogenous control using the following formula $2^{(Ct_{Gi} - Mean(Ct_{endogenouscontrol}))}$. For each biological replicate and target the expression level of all samples was divided by the one of the WT sample (arHIF[Col], Col-0 or WT) on control condition in order to set the level of expression of the WT sample to 1 (except in figure 3).

CATMA. The RNA of two pools of 12DAS seedlings grown on Man60 per genotype were extracted as described above and total RNA was checked for quality by nanochip analysis on the Agilent Bioanalyser and quantified using RiboGreen™ prior to microarray application. Transcript profiling was carried out on CATMA microarrays Version5 at URGV-Evry (INRA). Microarray hybridization, data analysis and CATdb database were described previously [56]. For the comparisons with Skirycz et al.'s analysis, the transcriptomic response of Col-0 treated with mannitol 25mM in expanding leaves was used. The overlap between the two sets that comprises genes that were differentially expressed in the same direction in the two analyses was tested using Genesect tool, a non-parametric randomization test available on the VirtualPlant1.3 website. The 20,250 TAIRv10 gene models common between the

CATMA.v5 and Affymetrix ATH1 arrays were used as the background list. Gene ontology analyses have been performed using Biomaps tool available on the VirtualPlant1.3 website that calculates p-values of over-representation using Fisher exact test with FDR correction [57].

GUS staining. Gus activity was performed on 12DAS seedlings (generation T3) grown in vitro under control and Man60 conditions without fixation. Infiltration with X-Glu buffer (sodium phosphate buffer 100 mM pH7, ferrocyanide 0.5 mM, ferricyanide 0.5 mM, X-Glu 8 mM diluted in DMSO) was carried out under vacuum for 10 min. The samples were then placed at 37°C overnight and discoloured the next day with 95% ethanol and the days after in 70% ethanol.

Phylogenetic analyses. A. thaliana members of the SD1 subfamily of RLKs were retrieved from Shiu et al. 2009 and as many SD1 *A. lyrata* homologs as possible were retrieved using Protein homologs search available on genes pages of Phytozome V9.0 website. Protein sequences of 34 *A. thaliana* RLK (32 SD1 members and two outgroup RLKs) and 43 *A. lyrata* RLKs (including two close homologs of *A. thaliana* outgroup) were aligned using the program MUSCLE implemented in SeaView v4.2 [58] and the obtained alignment was manually improved in GeneDoc v2.6.002 [59]. The final alignment restricted to the kinase domain (from amino acid 497 to amino acid 794 of EGM1) was used to align corresponding DNA coding sequences. To obtain the corresponding phylogeny, the best nucleotide substitution model fitting this alignment (GTR+I+G: GTR model with estimated proportion of invariable sites and Γ distribution) according to Akaike information criterion was determined using jModelTest v2.1.1. and used to run a Bayesian inference analysis using MrBayes v3.2.1 [60] with 1,000,000 generations and a burn-in of 5,000 generations. The resulting phylogeny was visualised in Treedyn [61].

Botrytis assays. All Botrytis cinerea infections were done as previously described using two previously characterized isolates ([62]). All plants were grown in a randomized complete block design and three leaves were taken from each plant. One for the control and one for each of the two isolates with a minimum of 10 leaves per infection. Tissue was harvested from the uninfected part of each leaf, the lesion and the lesion boundary (periphery) for each lesion, and analyzed for mannitol using a GC-TOF-MS based broad spectrum platform ([63]).

References

[1] Verslues, P. and Agarwal, M. and Katiyar-Agarwal, S. and Zhu, J. and Zhu, J., *Methods and concepts in quantifying resistance to drought, salt and freezing, abiotic stresses that affect plant water status.*, Plant J **45(4)**, 523-539 (2006).

[2] Skirycz, A. et al., *Developmental stage specificity and the role of mitochondrial metabolism in the response of arabidopsis leaves to prolonged mild osmotic stress.*, Plant Physiol **152(1)**, 226-244 (2010).

[3] Chan, Z. and Grumet, R. and Loescher, W., *Global gene expression analysis of transgenic, mannitol-producing, and salt-tolerant arabidopsis thaliana indicates widespread changes in abiotic and biotic stress-related genes.*, J Exp Bot , (2011).

[4] Stoop, J. and Williamson, J. and Mason Pharr, D., *Mannitol metabolism in plants: A method for coping with stress*, Trends in Plant Science **1(5)**, 139-144 (1996).

[5] Reinders, A. and Panshyshyn, J. and Ward, J., *Analysis of transport activity of arabidopsis sugar alcohol permease homolog AtPLT5.*, J Biol Chem **280(2)**, 1594-1602 (2005).

[6] Klepek, Y. and Geiger, D. and Stadler, R. and Klebl, F. and Landouar-Arsivaud, L. and Lemoine, R. and Hedrich, R. and Sauer, N., *Arabidopsis polyol transporter5, a new member of the monosaccharide transporter-like superfamily, mediates H+-symport of numerous substrates, including myo-inositol, glycerol, and ribose.*, Plant Cell **17(1)**, 204-218 (2005).

[7] Velez, H. and Glassbrook, N. and Daub, M., *Mannitol biosynthesis is required for plant pathogenicity by Alternaria alternata.*, FEMS Microbiol Lett **285(1)**, 122-129 (2008).

[8] Trontin, C. and Tisne, S. and Bach, L. and Loudet, O., *What does Arabidopsis natural variation teach us (and does not teach us) about adaptation in plants?*, Curr Opin Plant Biol **14(3)**, 225-231 (2011).

[9] Poormohammad Kiani, S. and Trontin, C. and Andreatta, M. and Simon, M. and Robert, T. and Salt, D. and Loudet, O., *Allelic heterogeneity and trade-off shape natural variation for response to soil micronutrient.*, PLoS Genet **8(7)**, e1002814 (2012).

[10] Masle, J. and Gilmore, S. and Farquhar, G., *The erecta gene regulates plant transpiration efficiency in arabidopsis.*, Nature **436(7052)**, 866-870 (2005).

[11] Shiu, S. and Bleecker, A., *Plant receptor-like kinase gene family: Diversity, function, and signaling.*, Sci STKE **2001(113)**, RE22 (2001).

[12] Lehti-Shiu, M.D. and Zou, C. and Hanada, K. and Shiu, S., *Evolutionary history and stress regulation of plant receptor-like kinase/pelle genes.*, Plant Physiol **150(1)**, 12-26 (2009).

[13] Ramachandraiah, G. and Chandra, N., *Sequence and structural determinants of mannose recognition.*, Proteins **39(4)**, 358-364 (2000).

[14] Wasano, N. and Ohgushi, A. and Ohba, M., *Mannose-specific lectin activity of parasporal proteins from a lepidoptera-specific bacillus thuringiensis strain.*, Curr Microbiol **46(1)**, 43-46 (2003).

[15] Tordai, H. and Banyai, L. and Patthy, L., *The pan module: The n-terminal domains of plasminogen and hepatocyte growth factor are homologous with the apple domains of the prekallikrein family and with a novel domain found in numerous nematode proteins.*, FEBS Lett **461(1-2)**, 63-67 (1999).

[16] Kilian, J. et al., *The atgenexpress global stress expression data set: Protocols, evaluation and model data analysis of UV-B light, drought and cold stress responses.*, Plant J **50(2)**, 347-363 (2007).

[17] Jennings, D. and Ehrenshaft, M. and Pharr, D. and Williamson, J., *Roles for mannitol and mannitol dehydrogenase in active oxygen-mediated plant defense.*, Proc Natl Acad Sci U S A **95(25)**, 15129-15133 (1998).

[18] Tarczynski, M. and Jensen, R. and Bohnert, H., *Stress protection of transgenic tobacco by production of the osmolyte mannitol.*, Science **259(5094)**, 508-510 (1993).

[19] Thomas, J. and Sepahi, M. and Arendall, B. and Bohnert, H., *Enhancement of seed germination in high salinity by engineering mannitol expression in Arabidopsis thaliana*, Plant, Cell & Environment **18(7)**, 801-806 (1995).

[20] Karakas, B. and Ozias-Akins2, P. and Stushnoff3, C. and Suefferheld, M. and Rieger, M., *Salinity and drought tolerance of mannitol-accumulating transgenic tobacco*, Plant, Cell & Environment **20(5)**, 609-616 (1997).

[21] Prabhavathi, V. and Yadav, J. and Kumar, P. and Rajam, M., *Abiotic stress tolerance in transgenic eggplant (Solanum melongena l.) by introduction of bacterial mannitol phosphodehydrogenase gene*, Molecular Breeding **9(2)**, 137-147 (2002).

[22] Abebe, T. and Guenzi, A. and Martin, B. and Cushman, J., *Tolerance of mannitol-accumulating transgenic wheat to water stress and salinity.*, Plant Physiol **131(4)**, 1748-1755 (2003).

[23] Zhifang G, L.W.H., *Expression of a celery mannose 6-phosphate reductase in* Arabidopsis thaliana *enhances salt tolerance and induces biosynthesis of both mannitol and a glucosyl-mannitol dimer*, Plant, Cell & Environment **26**, 275-283 (2003).

[24] Hu, L. and Lu, H. and Liu, Q. and Chen, X. and Jiang, X., *Overexpression of mtld gene in transgenic Populus tomentosa improves salt tolerance through accumulation of mannitol.*, Tree Physiol **25(10)**, 1273-1281 (2005).

[25] Fujita, M. et al., *Crosstalk between abiotic and biotic stress responses: A current view from the points of convergence in the stress signaling networks.*, Curr Opin Plant Biol **9(4)**, 436-442 (2006).

[26] Voegele, R. et al., *Possible roles for mannitol and mannitol dehydrogenase in the biotrophic plant pathogen Uromyces fabae.*, Plant Physiol **137(1)**, 190-198 (2005).

[27] Lewis, D.H. and Smith, D.C., *Sugar alcohols (polyols) in fungi and green plants: distribution, physiology and metabolism*, New Phytologist **66**, 143-184 (1967).

[28] Solomon, P. and Waters, O. and Oliver, R., *Decoding the mannitol enigma in filamentous fungi.*, Trends Microbiol **15(6)**, 257-262 (2007).

[29] Joosten, M.H.A.J. and Hendrickx, L.J.M. and De Wit, P.J.G.M., *Carbohydrate composition of apoplastic fluids isolated from tomato leaves inoculated with virulent or avirulent races of cladosporium fulvum (syn. Fulvia fulva)*, Netherlands Journal of Plant Pathology **96(2)**, 103-112 (1990).

[30] Dulermo, T. et al., *Dynamic carbon transfer during pathogenesis of sunflower by the necrotrophic fungus botrytis cinerea: From plant hexoses to mannitol.*, New Phytol **183(4)**, 1149-1162 (2009).

[31] Parker, D. and Beckmann, M. and Zubair, H. and Enot, D. and Caracuel-Rios, Z. and Overy, D. and Snowdon, S. and Talbot, N. and Draper, J., *Metabolomic analysis reveals a common pattern of metabolic re-programming during invasion of three host plant species by magnaporthe grisea.*, Plant J **59(5)**, 723-737 (2009).

[32] Smirnoff, N. and Cumbes, Q., *Hydroxyl radical scavenging activity of compatible solutes.*, Phytochemistry **28**, 1057-1060 (1989).

[33] Orthen, B., Popp, M. And Smirnoff, N., *Hydroxyl radical scavenging properties of cyclitols.*, Proc. R. Soc. Edinburgh Sect. B **102**, 269-272 (1994).

[34] Shen, B. and Jensen, R. and Bohnert, H., *Increased resistance to oxidative stress in transgenic plants by targeting mannitol biosynthesis to chloroplasts.*, Plant Physiol **113(4)**, 1177-1183 (1997).

[35] Shen, B. and Jensen, R. and Bohnert, H., *Mannitol protects against oxidation by hydroxyl radicals.*, Plant Physiol **115(2)**, 527-532 (1997).

[36] Chaturvedi, V. and Flynn, T. and Niehaus, W. and Wong, B., *Stress tolerance and pathogenic potential of a mannitol mutant of cryptococcus neoformans.*, Microbiology **142(Pt 4)**, 937-943 (1996).

[37] Chaturvedi, V. and Wong, B. and Newman, S., *Oxidative killing of cryptococcus neoformans by human neutrophils. Evidence that fungal mannitol protects by scavenging reactive oxygen intermediates.*, J Immunol **156(10)**, 3836-3840 (1996).

[38] Chaturvedi, V. and Bartiss, A. and Wong, B., *Expression of bacterial mtld in saccharomyces cerevisiae results in mannitol synthesis and protects a glycerol-defective mutant from high-salt and oxidative stress.*, J Bacteriol **179(1)**, 157-162 (1997).

[39] Torres, M., *Ros in biotic interactions.*, Physiol Plant **138(4)**, 414-429 (2010).

[40] Williamson, J. and Stoop, J. and Massel, M. and Conkling, M. and Pharr, D., *Sequence analysis of a mannitol dehydrogenase cdna from plants reveals a function for the pathogenesis-related protein ELI3.*, Proc Natl Acad Sci U S A **92(16)**, 7148-7152 (1995).

[41] Zamski, E. and Guo, W. and Yamamoto, Y. and Pharr, D. and Williamson, J., *Analysis of celery (apium graveolens) mannitol dehydrogenase (MTD) promoter regulation in arabidopsis suggests roles for mtd in key environmental and metabolic responses.*, Plant Mol Biol **47(5)**, 621-631 (2001).

[42] Jennings, D. and Daub, M. and Pharr, D. and Williamson, J., *Constitutive expression of a celery mannitol dehydrogenase in tobacco enhances resistance to the mannitol-secreting fungal pathogen alternaria alternata.*, Plant J **32(1)**, 41-49 (2002).

[43] Cheng, F. and Zamski, E. and Guo, W. and Pharr, D. and Williamson, J., *Salicylic acid stimulates secretion of the normally symplastic enzyme mannitol dehydrogenase: A possible defense against mannitol-secreting fungal pathogens.*, Planta , (2009).

[44] Becraft, P., *Receptor kinase signaling in plant development.*, Annu Rev Cell Dev Biol **18**, 163-192 (2002).

[45] Gish, L. and Clark, S., *The rlk/pelle family of kinases.*, Plant J **66(1)**, 117-127 (2011).

[46] Kim, H. et al., *An s-locus receptor-like kinase plays a*

role as a negative regulator in plant defense responses., Biochem Biophys Res Commun **381(3)**, 424-428 (2009).

[47] Pastuglia, M. et al., *Comparison of the expression patterns of two small gene families of s gene family receptor kinase genes during the defence response in brassica oleracea and arabidopsis thaliana.*, Gene **282(1-2)**, 215-225 (2002).

[48] Sanabria, N. and Van Heerden, H. and Dubery, I., *Molecular characterisation and regulation of a nicotiana tabacum s-domain receptor-like kinase gene induced during an early rapid response to lipopolysaccharides.*, Gene **501(1)**, 39-48 (2012).

[49] Kim, Y. and Oh, J. and Kim, K. and Uhm, J. and Lee, B., *Isolation and characterization of NgRLK1, a receptor-like kinase of nicotiana glutinosa that interacts with the elicitin of phytophthora capsici.*, Mol Biol Rep **37(2)**, 717-727 (2010).

[50] Chen, X. et al., *A B-lectin receptor kinase gene conferring rice blast resistance.*, Plant J **46(5)**, 794-804 (2006).

[51] Vlad, D. and Rappaport, F. and Simon, M. and Loudet, O., *Gene transposition causing natural variation for growth in arabidopsis thaliana.*, PLoS Genet **6(5)**, e1000945 (2010).

[52] Bouchabke, O. and Chang, F. and Simon, M. and Voisin, R. and Pelletier, G. and Durand-Tardif, M., *Natural variation in arabidopsis thaliana as a tool for highlighting differential drought responses.*, PLoS One **3(2)**, e1705 (2008).

[53] Simon, M. and Loudet, O. and Durand, S. and Berard, A. and Brunel, D. and Sennesal, F. and Durand-Tardif, M. and Pelletier, G. and Camilleri, C., *Quantitative trait loci mapping in five new large recombinant inbred line populations of Arabidopsis thaliana genotyped with consensus single-nucleotide polymorphism markers.*, Genetics **178(4)**, 2253-2264 (2008).

[54] Wesley, S. and Helliwell, C. and Smith, N. and Wang, M. and Rouse, D. and Liu, Q. and Gooding, P. and Singh, S. and Abbott, D. and Stoutjesdijk, P. and Robinson, S. and Gleave, A. and Green, A. and Waterhouse, P., *Construct design for efficient, effective and high-throughput gene silencing in plants.*, Plant J **27(6)**, 581-590 (2001).

[55] Hellens, R. and Edwards, E. and Leyland, N. and Bean, S. and Mullineaux, P., *Pgreen: A versatile and flexible binary ti vector for agrobacterium-mediated plant transformation.*, Plant Mol Biol **42(6)**, 819-832 (2000).

[56] Gagnot, S. and Tamby, J. and Martin-Magniette, M.L. and Bitton, F. and Taconnat, L. and Balzergue, S. and Aubourg, S. and Renou, J. and Lecharny, A. and Brunaud, V., *Catdb: A public access to arabidopsis transcriptome data from the urgv-catma platform.*, Nucleic Acids Res **36(Database issue)**, D986-90 (2008).

[57] Katari, M. and Nowicki, S. and Aceituno, F. and Nero, D. and Kelfer, J. and Thompson, L. and Cabello, J. and Davidson, R. and Goldberg, A. and Shasha, D. and Coruzzi, G. and Gutierrez, R., *Virtualplant: A software platform to support systems biology research.*, Plant Physiol **152(2)**, 500-515 (2010).

[58] Gouy, M. and Guindon, S. and Gascuel, O., *Seaview version 4 : A multiplatform graphical user interface for sequence alignment and phylogenetic tree building.*, Mol Biol Evol , (2009).

[59] Nicholas, K. and Nicholas, H. and Deerfield, D., *Genedoc: Analysis and visualization of genetic variation*, Embnet News **4**, 14 (1997).

[60] Huelsenbeck, J. and Ronquist, F., *Mrbayes: Bayesian inference of phylogenetic trees*, Bioinformatics **17(8)(8)**, 754-755 (2001).

[61] Chevenet, F. and Brun, C. and Banuls, A. and Jacq, B. and Christen, R., *Treedyn: Towards dynamic graphics and annotations for analyses of trees.*, BMC Bioinformatics **7**, 439 (2006).

[62] Rowe, H. and Walley, J. and Corwin, J. and Chan, E. and Dehesh, K. and Kliebenstein, D., *Deficiencies in jasmonate-mediated plant defense reveal quantitative variation in botrytis cinerea pathogenesis.*, PLoS Pathog **6(4)**, e1000861 (2010).

[63] Fiehn, O. and Wohlgemuth, G. and Scholz, M. and Kind, T. and Lee Do, Y. and Lu, Y. and Moon, S. and Nikolau, B., *Quality control for plant metabolomics: Reporting msi-compliant studies.*, Plant J **53(4)**, 691-704 (2008).

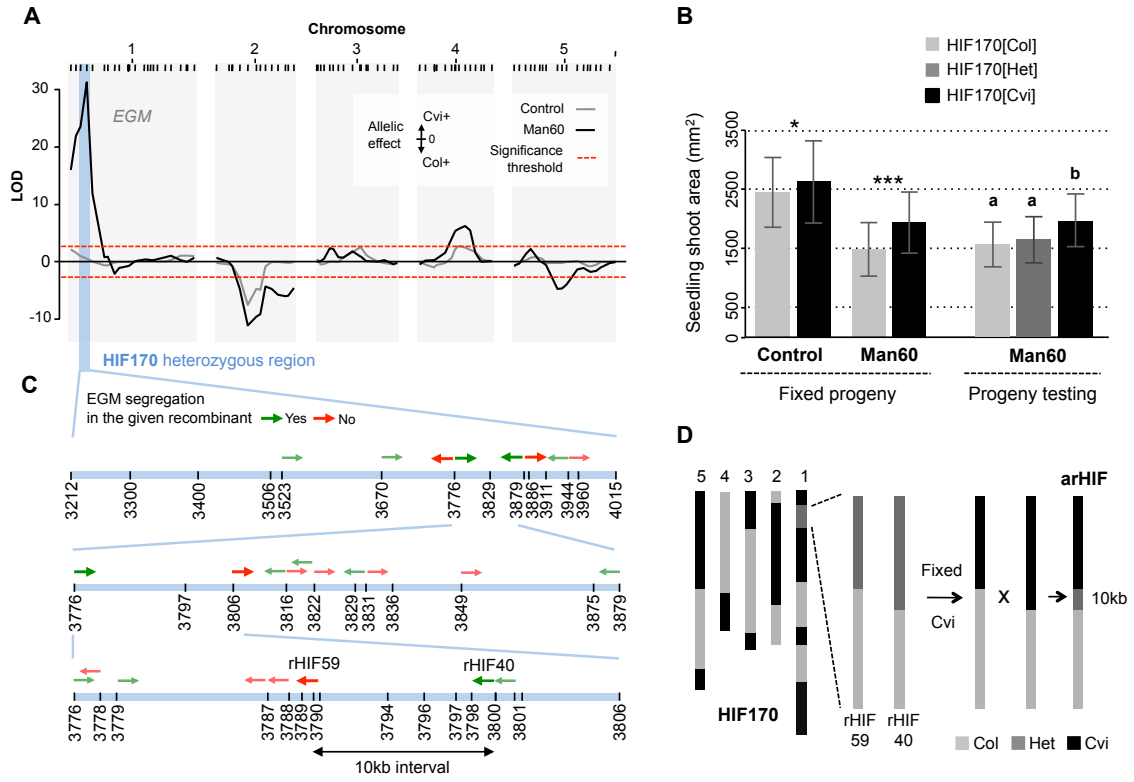


FIG. S1. **EGM QTL mapping and cloning.** A. LOD score curves obtained from multiple QTL mapping analyses implemented in R for shoot area under control and Man60 conditions in the Cvi-0xCol-0 RIL set. Positive LOD scores correspond to regions where the Cvi allele has a positive effect on the phenotype whereas negative LOD scores correspond to regions of Col-positive effects. B. EGM QTL confirmation in the heterogeneous inbred family HIF170 segregating for the top of chromosome 1. In fixed progeny experiments, the descendants of the two fixed HIF170 (HIF170[Col] and HIF170[Cvi]) were phenotyped on Control and Man60 condition. In progeny testing experiment, the descendants of the heterozygous RIL170 were directly genotyped and phenotyped on Man60, reducing possible seed stock- and maternal-effects. Error bars represent standard errors obtained from the phenotyping of at least 40 individuals. Stars represent significant genotypic effects obtained for each media with a Student t-test (* $0.01 < p < 0.05$; *** $p < 0.001$) and letters indicate significantly different genotypes (Tukeys HSD, $p < 0.05$). C. The analysis of EGM segregation in a series of recombinants issued from HIF170[Het] reduced the QTL candidate interval to 10kb. One arrow represents the position of a recombination point from an individual recombinant: for each, the back of the arrow indicates the fixed (homozygous) region whereas the head points toward the heterozygous region. A green (red) arrow means that EGM was found segregating (or not) in the recombinant. D. Generation of advanced recombined HIF from rHIF59 and rHIF40.

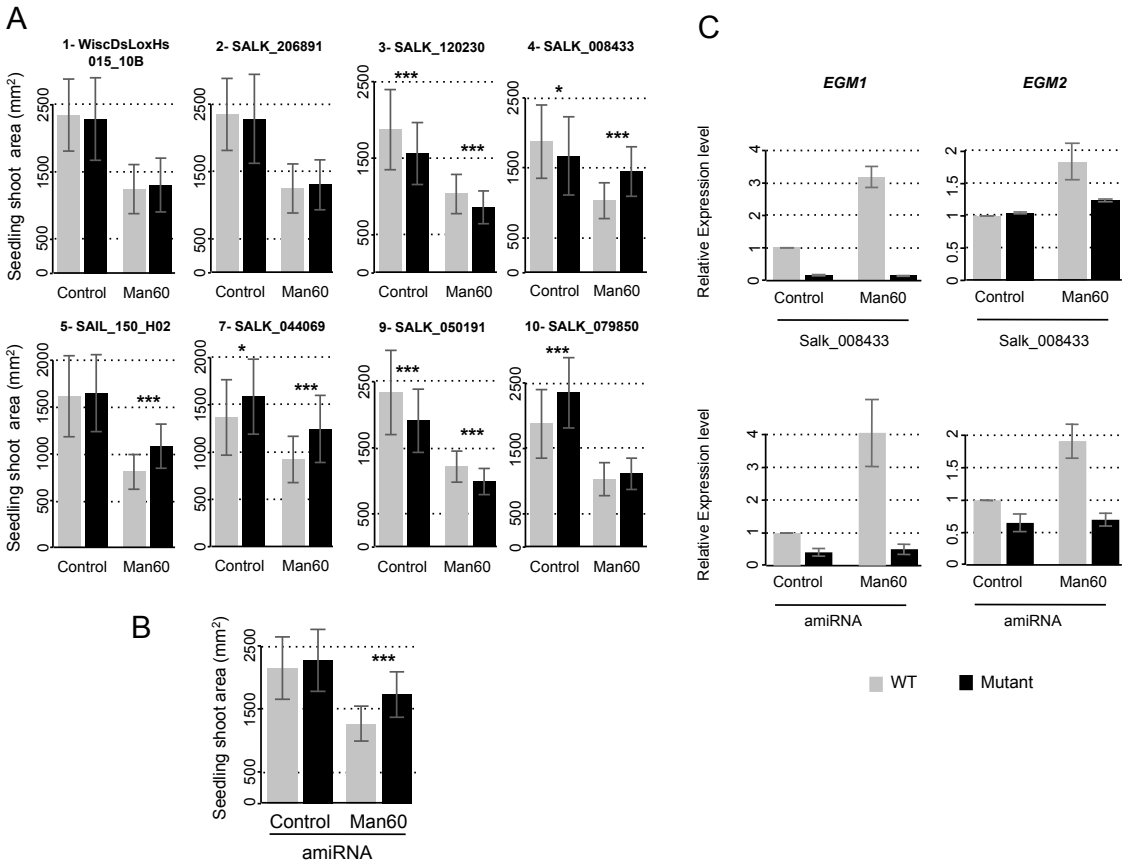


FIG. S2. **T-DNA mutant analysis revealed that *EGM1* and *EGM2* are strong candidates for the EGM QTL.** A. Phenotyping of several T-DNA mutants on control and Man60 conditions. Approximate position of the T-DNAs is indicated on Figure 1B. Error bars represent standard errors obtained from the phenotyping of at least 30 plants and a second biological replicate gave similar results. B. Phenotyping of an amiRNA line on control and Man60 conditions. Error bars represent standard deviations obtained from the phenotyping of at least 30 plants for two individual transformants and a second biological replicate gave similar results. Stars represent the significant genotypic effects obtained for each media with a Student t-test (* $0.01 < p < 0.05$; *** $p < 0.001$). C. Relative *EGM1* and *EGM2* expression levels in 12 DAS seedlings of SALK_008433, amiRNA and their respective WT lines grown under control and Man60 conditions. Error bars represent the standard deviation observed in two biological replicates obtained from a single (SALK_008433) or two (amiRNA) independent experiments.

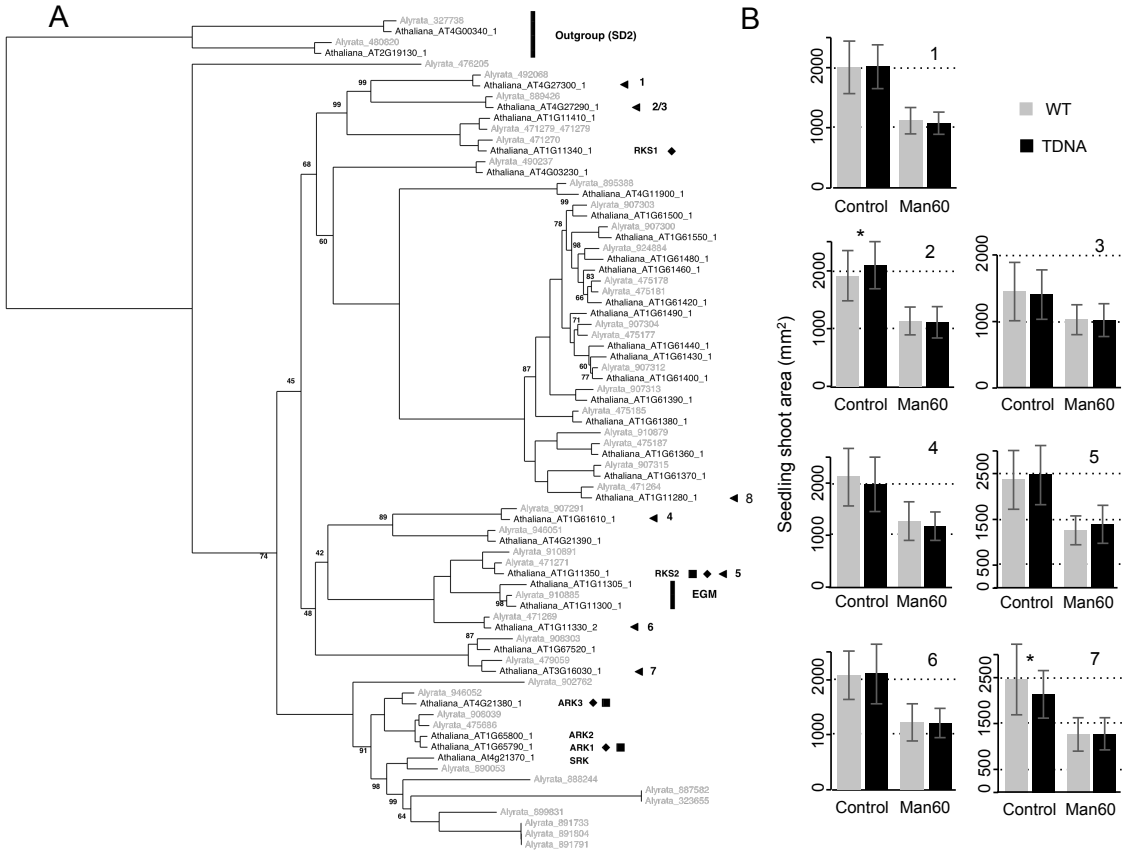


FIG. S3. **Evolution of the SD1 subfamily.** A. Phylogeny of the members of the SD1 subfamily in *A. thaliana* and *A. lyrata* based on an analysis with MrBayes of the nucleotide residues encoding the kinase domain of the RLKs. The name of each sequence analysed includes the name of the species followed by the transcript name retrieved from PhytozomeV9 gene pages. The number on each node corresponds to clade support values in %. For more clarity, the support values of 100% have been omitted. *At4g00340* was used as outgroup in the analysis. The names of the genes already published are indicated in bold. Diamonds indicate genes upregulated by salicylic acid treatment and squares genes upregulated after bacterial infection or wounding. B. Phenotyping of several T-DNA mutants in the genes indicated by triangles on the phylogenetic tree (A) under control and Man60 conditions. 1: SALK_145131; 2: SAIL_629_C11; 3: SALK_067606; 4: SALK_076477; 5: SALK_099776; 6: SALK_143489; 7: SAIL_1212_G06; 8: see Fig. 1B and S2A (*At1g11280*). Error bars represent the standard deviation obtained from the phenotyping of at least 30 plants and a second biological replicate gave similar results. Stars represent the significant genotypic effects obtained for each media with a Student t-test (* $0.01 < p < 0.05$).

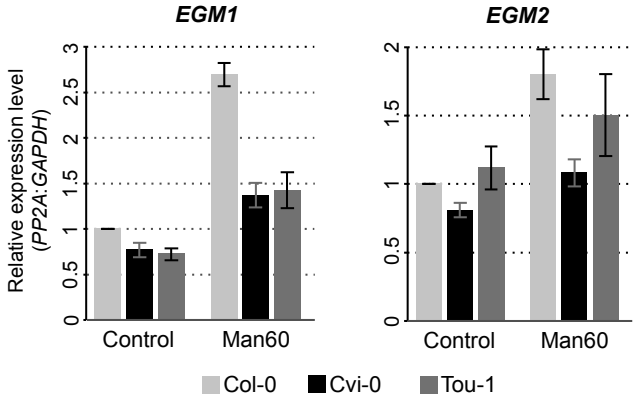


FIG. S4. **Relative *EGM1* and *EGM2* expression levels in Col-0, Cvi-0 and Tou-1 accessions.** 12 DAS plants grown under control and Man60 conditions were used to estimate expression levels. Error bars represent the standard deviation observed among two biological replicates obtained from a single experiment.

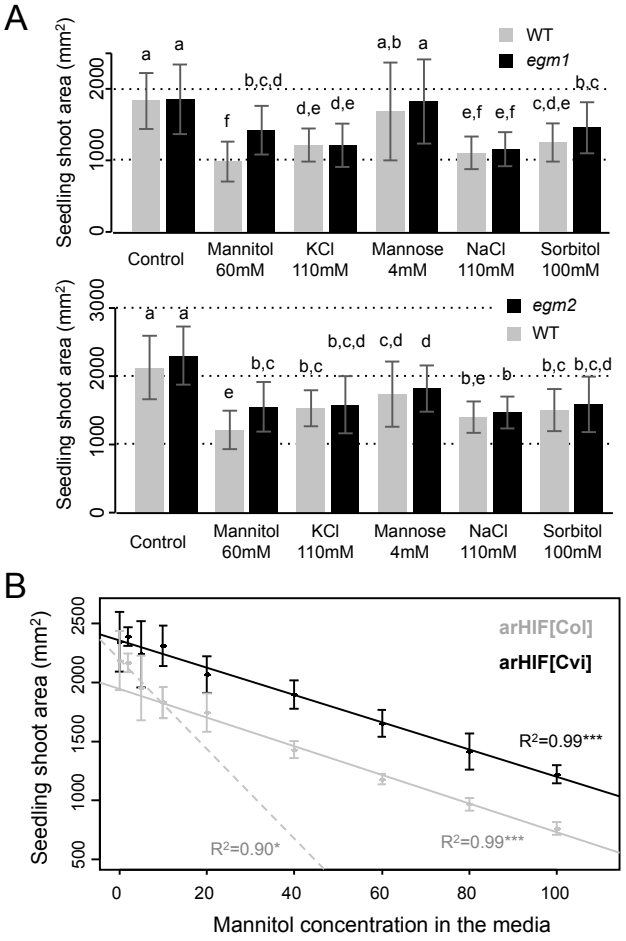


FIG. S5. **Additional data regarding the unlikely role of *EGM1* and *EGM2* in abiotic stress response.** A. Phenotyping of 12DAS seedlings of *egm1* and *egm2* mutants grown under different osmotic constraints. Different letters indicate significantly different groups based on a post-hoc Tukeys HSD test ($p < 0.05$). Error bars represent standard deviation obtained from the phenotyping of at least 30 plants and a second biological replicate gave similar results. B. Phenotyping of 12 DAS seedlings of the arHIF[Col] and arHIF[Cvi] grown on media containing different concentrations of mannitol (0 to 100mM). For the arHIF[Cvi] the regression was performed between the concentration 0 to 100mM. For the arHIF[Col] the regressions were performed between 0 to 10mM (dashed line) and 10 to 100mM (solid line). Stars indicate the significance of the effect of mannitol concentration on seedling growth (* $p < 0.05$, *** $p < 0.001$). Adjusted r-square values are also indicated. Error bars represent standard deviation obtained from the phenotyping of at least 3 pools of 20 plants in two independent experiments.

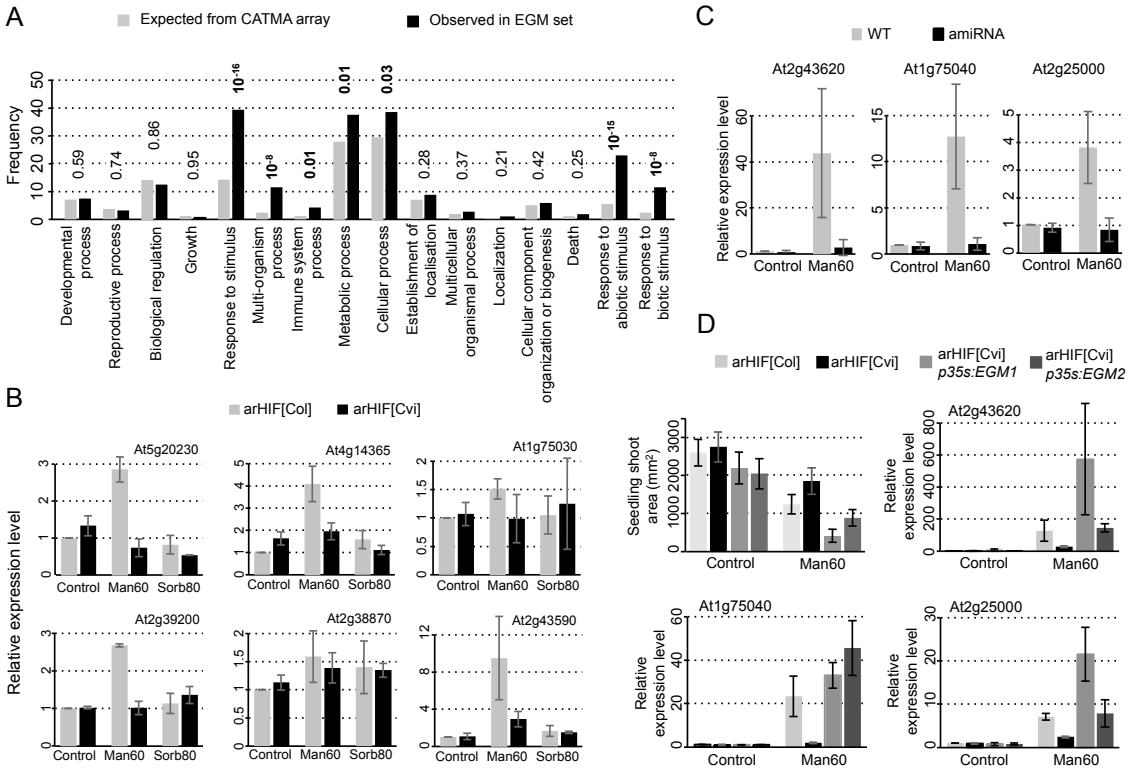


FIG. S6. **CATMA array results and additional qPCR validation.** A. Comparison of the transcriptomes of the arHIF[Col] and arHIF[Cvi] 12 DAS seedlings grown under Man60 using CATMA arrays identified 221 genes differentially expressed between the two genotypes ('EGM set'). The frequencies of biological process GO categories expected from the CATMA array and observed in 'EGM set' were compared using Fisher exact test with FDR correction. 3 to 6 genes differentially expressed between the arHIF[Col] and arHIF[Cvi] on Man60 in our transcriptomic analysis, and known to be induced by mannitol treatment (Skirycz et al. 2011) were analysed by quantitative RT-PCR in 12 DAS seedlings of the arHIFs (B), amiRNA (C) and over-expressor (D) lines on control, Man60 and/or Sorb80 conditions. Error bars represent the standard deviation observed in two biological replicates obtained from two independent experiments. In panel D, the phenotype data of over-expressors on control and Man60 condition was also plotted and the level of expression of *EGM1* and *EGM2* in those lines is available in Figure 2. Error bars represent standard deviation obtained from the phenotyping of at least 30 plants and a second biological replicate gave similar results.

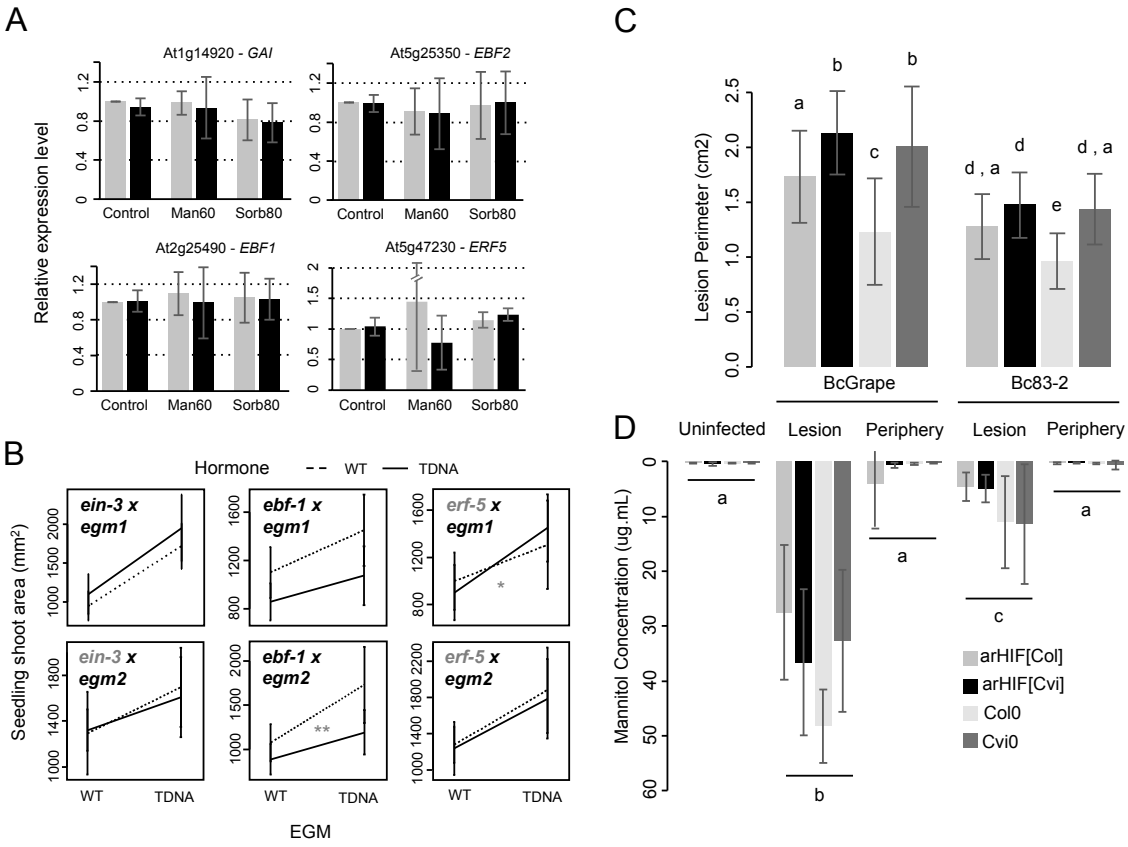
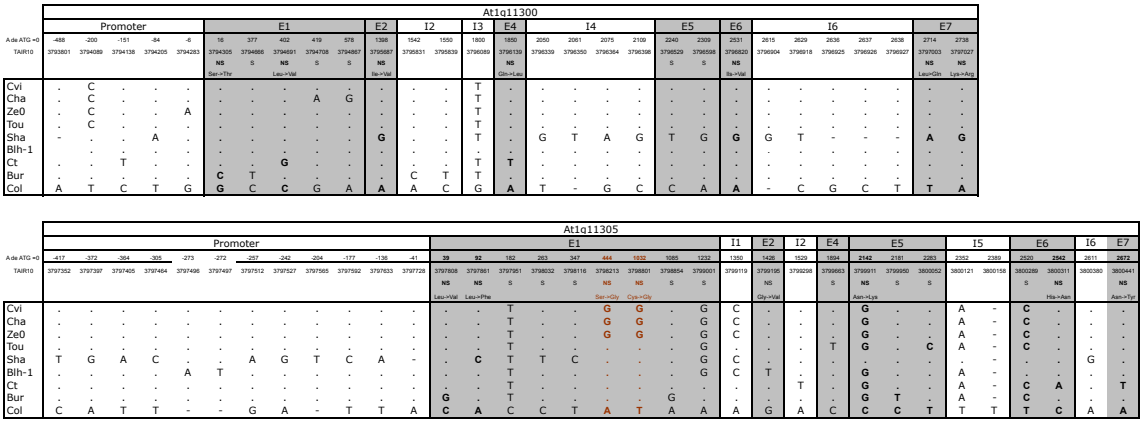


FIG. S7. **The role *EGM1* and *EGM2* in biotic stress response.** A. Four genes differentially expressed between the arHIF[Col] and arHIF[Cvi] on Man60 in our transcriptome analysis and involved in ethylene signaling pathway were analysed by quantitative RT-PCR in 12 DAS seedlings of the arHIFs in control, Man60 and Sorb80 conditions. Note that the expression level of *ERF5* in the arHIF[Col] was higher than the one of the arHIF[Cvi] on Man60 in two independent replicates but that the level on the other media varied between the two experiments. B. *egm1* and *egm2* mutants were crossed with 3 different mutants in ethylene signaling pathways (and their respective WT); single and double mutants were phenotyped for seedling growth on Man60 12 DAS. For each interaction plot, the absence of effect of egm mutation or homonal mutation is indicated in grey in the name of the cross (ANOVA; $p > 0.05$). The significance of the effect of the interaction between egm and hormonal mutations is indicated by stars (ANOVA; *: $p < 0.05$, ** $p < 0.01$). Grey stars are interactions that were not confirmed in a second experiment. Error bars represent standard deviation obtained from the phenotyping of at least 30 plants. C. Mean perimeter of necrotic lesions formed by *B. cinerea* isolates BcGrape or Bc83-2 on arHIF[Col], arHIF[Cvi], Col-0 and Cvi-0 plants 72 hours post-inoculation. Different letters indicate significantly different groups based on a post-hoc Tukeys HSD test ($p < 0.05$). D. Mannitol concentrations observed within or at the periphery of necrotic lesions formed by *B. cinerea* isolates BcGrape or Bc83-2 on arHIF[Col], arHIF[Cvi], Col-0 and Cvi-0 plants 72 hours post-inoculation. Different letters indicate significantly different groups based on a post-hoc Kruskal-Wallis test ($p < 0.05$).



TAB. S1. **Polymorphisms observed in EGM1 and EGM2 in the accessions used for F2 crosses.**

EGM		Log2 fold change (arHIF[Col]-arHIF[Cvi])	Skirycz	Common to EGM and Skirycz
			Log2 fold change (E25-E0)	NA: not in array - 0: not common - 1 common
At2g42540	COR15, COR15A, cold-regulated 15a	-1.12	0.88	0
At1g62305	Core-2/I-branching beta-1,6-N-acetylglucosaminyltransferase family protein	-1.06	NA	NA
At1g54040	ESP, ESR, TASTY, epithiospecifier protein	-0.98	0	0
At1g06080	ADS1, delta 9 desaturase 1	-0.88	0	0
At5g24420	PGL5, 6-phosphogluconolactonase 5	-0.83	0	0
At3g28300	AT14A, Protein of unknown function (DUF677)	-0.81	NA	NA
At2g07739	Ycf1 protein	-0.79	NA	NA
At4g30450	glycine-rich protein	-0.79	0	0
At5g40450	unknown protein: FUNCTIONS IN: molecular_function unknown: INVOLVED IN: biological	-0.79	0	0
At1g13930	Involved in response to salt stress. Knockout mutants are hypersensitive to salt stress.	-0.78	0	0
At1g44900	ATMCM2, MCM2, minichromosome maintenance (MCM2/3/5) family protein	-0.78	0	0
At3g28270	Protein of unknown function (DUF677)	-0.78	0	0
At4g16146	cAMP-regulated phosphoprotein 19-related protein	-0.78	0	0
At1g05010	ACO4, EAT1, EFE, ethylene-forming enzyme	-0.75	0	0
At1g74890	ARR15, response regulator 15	-0.75	-0.98	1
At4g37980	ATCAD7, CAD7, ELI3, ELI3-1, elicitor-activated gene 3-1	-0.75	0	0
At5g36910	THI2.2, thionin 2.2	-0.74	-0.66	1
At2g27385	Pollen Ole e 1 allergen and extensin family protein	-0.72	NA	NA
At3g05880	RCI2A, Low temperature and salt responsive protein family	-0.71	0	0
At5g11420	Protein of unknown function, DUF642	-0.71	0	0
At5g62040	PEBP (phosphatidylethanolamine-binding protein) family protein	-0.71	0.71	0
At1g69680	Mog1/PsbP/DUF1795-like photosystem II reaction center PsbP family protein	-0.68	0	0
At1g43690	ubiquitin interaction motif-containing protein	0.68	0	0
At1g70980	SYNC3, Class II aminoacyl-tRNA and biotin synthetases superfamily protein	0.68	NA	NA
At2g05380	GRP3S, glycine-rich protein 3 short isoform	0.68	1.38	1
At2g46500	ATPI4K GAMMA 4, PI4K GAMMA 4, UBDK GAMMA 4, phosphoinositide 4-kinase gamma 4	0.68	0	0
At3g60750	Transketolase	0.68	0	0
At1g56070	LOS1, Ribosomal protein S5/Elongation factor G/III/V family protein	0.69	0	0
At2g04030	AtHsp90.5, CR88, EMB1956, Hsp88.1, HSP90.5, Chaperone protein htpG family protein	0.69	0	0
At2g26140	ftsh4, FTSH protease 4	0.69	0	0
At2g42810	PAPP5, PP5, PP5.2, protein phosphatase 5.2	0.69	0	0
At3g03780	ATMS2, MS2, methionine synthase 2	0.69	NA	NA
At4g25730	Ftsi-like methyltransferase family protein	0.69	0	0
At5g47230	**ATERF-5, ATERF5, ERF5, ethylene responsive element binding factor 5**	0.69	1.06	1
At2g24600	Ankyrin repeat family protein	0.7	1.38	1
At2g31560	Protein of unknown function (DUF1685)	0.7	NA	NA
At2g34750	RNA polymerase I specific transcription initiation factor RRN3 protein	0.7	0	0
At2g45290	Transketolase	0.7	0	0
At2g45970	CYP86A8, LCR, cytochrome P450, family 86, subfamily A, polypeptide 8	0.7	0	0
At4g23180	CRK10, RLK4, cysteine-rich RLK (RECEPTOR-like protein kinase) 10	0.7	1.84	1
At5g22740	ATCSLA02, ATCSLA2, CSLA02, CSLA2, cellulose synthase-like A02	0.7	0	0
At5g25100	Endomembrane protein 70 protein family	0.7	0	0
At5g42530	unknown protein: FUNCTIONS IN: molecular_function unknown: INVOLVED IN: biological	0.7	NA	NA
At1g36360	transposable element gene	0.71	NA	NA
At2g01670	atnudt17, NUDT17, nudix hydrolase homolog 17	0.71	0.6	1
At2g34250	SecY protein transport family protein	0.71	0	0
At3g07270	GTP cyclohydrolase I	0.71	0	0
At5g55750	hydroxyproline-rich glycoprotein family protein	0.71	0	0
At1g27090	glycine-rich protein	0.72	0	0
At1g68840	EDF2, RAP2.8, RAV2, TEM2, related to ABI3/VP1 2	0.72	0.59	1
At1g80300	ATNTT1, NTT1, nucleotide transporter 1	0.72	NA	NA
At4g02715	unknown protein: Has 30201 Blast hits to 17322 proteins in 780 species: Archae - 12: Bac	0.72	NA	NA
At5g17990	pat1, TRP1, tryptophan biosynthesis 1	0.72	0	0
At1g09420	G6PD4, glucose-6-phosphate dehydrogenase 4	0.73	0.49	1
At1g09970	LRR XI-23, RLK7, Leucine-rich receptor-like protein kinase family protein	0.73	1.19	1
At1g29350	Kinase-related protein of unknown function (DUF1296)	0.73	NA	NA
At3g10985	ATWI-12, SAG20, WI12, senescence associated gene 20	0.73	0	0
At3g11710	ATKRS-1, lysyl-tRNA synthetase 1	0.73	0	0
At3g12580	ATHSP70, HSP70, heat shock protein 70	0.73	0	0
At3g16460	Mannose-binding lectin superfamily protein	0.73	1.61	1
At3g55980	ATSZF1, SZF1, salt-inducible zinc finger 1	0.73	1.72	1
At1g21750	ATPDIS, ATPDIL1-1, PDIS, PDIL1-1, PDI-like 1-1	0.74	0.54	1
At1g49630	ATPREP2, PREP2, presequence protease 2	0.74	0	0
At2g38860	YLS5, Class I glutamine amidotransferase-like superfamily protein	0.74	1.91	1
At2g42520	P-loop containing nucleoside triphosphate hydrolases superfamily protein	0.74	0	0
At5g20250	DIN10, Raffinose synthase family protein	0.74	-1.21	0
At5g25340	Ubiquitin-like superfamily protein	0.74	0	0
At3g16000	MFP1, MAR binding filament-like protein 1	0.75	0	0
At3g22060	Receptor-like protein kinase-related family protein	0.75	1.22	1
At3g28930	AIG2, AIG2-like (avirulence induced gene) family protein	0.75	0.85	1
At4g33220	ATPME44, PME44, pectin methylesterase 44	0.75	0.93	1
At5g35160	Endomembrane protein 70 protein family	0.75	0	0
At2g25510	unknown protein: Has 2 Blast hits to 2 proteins in 1 species: Archae - 0: Bacteria - 0: Met	0.76	0	0
At4g00370	ANTR2, PHT4;4, Major facilitator superfamily protein	0.76	0	0
At4g19210	ATRLI2, RLI2, RNAse l inhibitor protein 2	0.76	0	0
At4g22900	Protein of unknown function (DUF1191)	0.76	NA	NA
At2g05530	Glycine-rich protein family	0.77	0	0
At3g02450	cell division protein ftsH, putative	0.77	0	0
At3g07390	AIR12, auxin-responsive family protein	0.77	0	0
At4g00360	ATT1, CYP86A2, cytochrome P450, family 86, subfamily A, polypeptide 2	0.77	NA	NA
At4g24190	AtHsp90-7, AtHsp90.7, HSP90.7, SHD, Chaperone protein htpG family protein	0.77	0	0
At4g27300	S-locus lectin protein kinase family protein	0.77	2.41	1
At4g30910	Cytosol aminopeptidase family protein	0.77	NA	NA
At5g18170	GDH1, glutamate dehydrogenase 1	0.77	1.64	1
At1g14920	**GAI, RGA2, GRAS family transcription factor family protein**	0.78	0	0
At1g15500	ATNTT2, TLC ATP/ADP transporter	0.78	NA	NA
At1g30730	FAD-binding Berberine family protein	0.78	NA	NA
At1g72930	TIR, toll/interleukin-1 receptor-like	0.78	NA	NA
At4g23850	LACS4, AMP-dependent synthetase and ligase family protein	0.78	0	0
At5g42940	RING/U-box superfamily protein	0.78	0	0
At1g22890	unknown protein: FUNCTIONS IN: molecular_function unknown: INVOLVED IN: biological	0.79	0	0
At1g64370	unknown protein: Has 773 Blast hits to 375 proteins in 118 species: Archae - 0: Bacteria -	0.79	0	0
At3g50480	HR4, homolog of RPW8 4	0.79	1.68	1
At4g02380	AILEA5, SAG21, senescence-associated gene 21	0.79	3.22	1
At5g13400	Major facilitator superfamily protein	0.79	0	0
At5g22520	unknown protein: BEST Arabidopsis thaliana protein match is: unknown protein (TAIR:AT5	0.79	NA	NA
At5g28540	BIP1, heat shock protein 70 (Hsp 70) family protein	0.79	NA	NA
At2g16500	ADC1, ARGDC, ARGDC1, SPE1, arginine decarboxylase 1	0.8	-0.4	0
At3g26200	CYP71B22, cytochrome P450, family 71, subfamily B, polypeptide 22	0.8	0	0
At4g34110	ATPAB2, PAB2, PABP2, poly(A) binding protein 2	0.8	0	0
At5g26360	TCP-1/cpn60 chaperonin family protein	0.8	0	0
At5g45110	ATNPR3, NPR3, NPR1-like protein 3	0.8	0	0
At1g06290	ACX3, ATACX3, acyl-CoA oxidase 3	0.81	NA	NA
At1g13110	CYP71B7, cytochrome P450, family 71 subfamily B, polypeptide 7	0.81	0.91	1
At1g67280	Glyoxalase/Bleomycin resistance protein/Dioxygenase superfamily protein	0.81	0	0
At3g59970	MTHFR1, methylenetetrahydrofolate reductase 1	0.81	NA	NA
At4g38550	Arabidopsis phospholipase-like protein (PEARL) 4 family	0.81	2.42	1
At5g24160	SQE6, squalene monoxygenase 6	0.81	0.71	1
At1g65540	LETM1-like protein	0.82	0	0
At2g26560	PLA IIA, PLA2A, PLP2, PLP2, phospholipase A 2A	0.82	2.01	1

qPCR confirmed
qPCR non confirmed
ethylene responsive
response to biotic stimulus

EGM		Log2 fold change (arHIF[Col]-arHIF[Cvi])	Skirycz	Common to EGM and Skirycz
			Log2 fold change (E25-E0)	NA: not in array - 0: not common - 1 common
At2g36580	Pyruvate kinase family protein	0.82	NA	NA
At2g38470	**ATWRKY33, WRKY33, WRKY DNA-binding protein 33**	**0.82**	**2.75**	**1**
At2g39400	alpha/beta-Hydrolases superfamily protein	0.82	2.14	1
At3g46970	ATPHS2, PHS2, alpha-glucan phosphorylase 2	0.82	0	0
At4g12650	Endomembrane protein 70 protein family	0.83	0	0
At5g09590	HSC70-5, MTHSC70-2, mitochondrial HSO70 2	0.83	0	0
At1g80530	Major facilitator superfamily protein	0.84	0.59	1
At2g26250	FDH, KCS10, 3-ketoacyl-CoA synthase 10	0.84	0	0
At5g17920	ATCIMS, ATMETS, ATMS1, Cobalamin-independent synthase family protein	0.84	NA	NA
At1g64450	Glycine-rich protein family	0.85	0	0
At1g66970	SVL2, SHV3-like 2	0.85	0	0
At1g33610	Leucine-rich repeat (LRR) family protein	0.86	0	0
At4g33050	EDA39, calmodulin-binding family protein	0.86	1.88	1
At1g78080	RAP2.4, related to AP2 4	0.87	0	0
At2g28000	CH-CPN60A, CPN60A, SLP, chaperonin-60alpha	0.87	0	0
At3g18060	transducin family protein / WD-40 repeat family protein	0.87	0	0
At5g48380	BIR1, BAK1-interacting receptor-like kinase 1	0.87	1.34	1
At2g18660	PNP-A, plant natriuretic peptide A	0.88	4.64	1
At2g27050	AtEIL1, EIL1, ETHYLENE-INSENSITIVE3-like 1	0.88	0	0
At2g38870	**Serine protease inhibitor, potato inhibitor I-type family protein**	0.88	0.81	1
At4g36670	Major facilitator superfamily protein	0.88	0	0
At1g14870	PCR2, PLANT CADMIUM RESISTANCE 2	0.89	NA	NA
At2g39200	**ATMLO12, MLO12, Seven transmembrane MLO family protein**	0.89	3.08	1
At4g18250	receptor serine/threonine kinase, putative	0.89	0	0
At5g56030	AtHsp90.2, ERD8, HSP81-2, HSP90.2, heat shock protein 81-2	0.89	0	0
At3g50950	ZAR1, HOP2-ACTIVATED RESISTANCE 1	0.9	NA	NA
At2g24180	CYP718, cytochrome p450 718	0.91	0.94	1
At4g37910	mtHsc70-1, mitochondrial heat shock protein 70-1	0.91	0	0
At5g14120	Major facilitator superfamily protein	0.91	0	0
At4g03420	Protein of unknown function (DUF789)	0.92	NA	NA
At5g01870	Bifunctional inhibitor/lipid-transfer protein/seed storage 2S albumin superfamily protein	0.92	3.63	1
At1g49570	NQR, ARP protein (REF)	0.93	0	0
At1g55280	Lipase/lipooxygenase, PLAT/LH2 family protein	0.93	0	0
At3g02470	SAMDC, S-adenosylmethionine decarboxylase	0.93	0	0
At3g13750	BGAL1, BGAL1, beta galactosidase 1	0.93	-0.55	0
At4g00720	ASKTHETA, ATSK32, SK32, shaggy-like protein kinase 32	0.93	0	0
At4g11230	Riboflavin synthase-like superfamily protein	0.93	0	0
At4g22710	CYP706A2, cytochrome P450, family 706, subfamily A, polypeptide 2	0.93	NA	NA
At1g05300	ZIP5, zinc transporter 5 precursor	0.94	1.53	1
At1g37130	ATNR2, B29, CHL3, NIA2, NIA2-1, NR, NR2, nitrate reductase 2	0.95	0	0
At2g05520	ATGRP-3, ATGRP3, GRP-3, GRP3, glycine-rich protein 3	0.95	1.34	1
At3g56050	Protein kinase family protein	0.95	0	0
At5g25610	ATRD22, RD22, BURP domain-containing protein	0.95	1.38	1
At5g25350	**EBF2, EIN3-binding F box protein 2**	**0.96**	**0**	**0**
At5g49720	ATGHPA1, DEC, GHPA1, IRX2, KOR, KOR1, RSW2, TSD1, glycosyl hydrolase 9A1	0.96	0	0
At5g04590	SIR, sulfite reductase	0.97	0	0
At1g08510	FAT8, fatty acyl-ACP thioesterases B	0.98	0	0
At1g30100	ATNCED6, NCED5, nine-cis-epoxycarotenoid dioxygenase 5	0.98	0	0
At2g25000	**ATWRKY60, WRKY60, WRKY DNA-binding protein 60**	0.99	1.43	1
At3g31660	CONTAINS InterPro DOMAIN/s: Bystin (InterPro:IPR007955): Has 475 Blast hits to 467 proteins	1	0	0
At2g28630	KCS12, 3-ketoacyl-CoA synthase 12	1	1.02	1
At3g07700	Protein kinase superfamily protein	1	0	0
At1g30290	Tetratricopeptide repeat (TPR)-like superfamily protein	1.01	0	0
At1g18390	Protein kinase superfamily protein	1.04	1.53	1
At1g65845	unknown protein: FUNCTIONS IN: molecular_function unknown: INVOLVED IN: biological_proce	1.04	NA	NA
At2g16660	Major facilitator superfamily protein	1.04	0.59	1
At4g23100	ATECS1, CAD2, GSH1, GSHA, PAD2, RML1, glutamate-cysteine ligase	1.04	0	0
At3g20460	Major facilitator superfamily protein	1.05	0	0
At1g75030	**ATLP-3, TLP-3, thaumatin-like protein 3**	1.06	0.51	1
At2g26440	Plant invertase/pectin methylesterase inhibitor superfamily	1.06	1.24	1
At3g01290	SPFH/Band 7/PHB domain-containing membrane-associated protein family	1.06	1.94	1
At5g24110	ATWRKY30, WRKY30, WRKY DNA-binding protein 30	1.06	3.85	1
At1g22710	ATSUC2, SUC2, SUT1, sucrose-proton symporter 2	1.08	0	0
At1g26110	DCP5, decapping 5	1.08	0	0
At1g74960	ATKAS2, FAB1, KAS2, fatty acid biosynthesis 1	1.09	0	0
At2g31460	Domain of unknown function (DUF313)	1.09	0	0
At3g30775	AT-POX, ATPDH, ATPOX, ERD5, PRO1, PRODH, Methylenetetrahydrofolate reductase family prot	1.1	0	0
At3g47540	Chitinase family protein	1.1	0	0
At5g08610	P-loop containing nucleoside triphosphate hydrolases superfamily protein	1.11	0	0
At2g25490	**EBF1, FBL6, EIN3-binding F box protein 1**	**1.12**	**0.29**	**1**
At3g19930	ATSTP4, STP4, sugar transporter 4	1.13	1.45	1
At2g31880	EVR, SOBIR1, Leucine-rich repeat protein kinase family protein	1.14	NA	NA
At2g47240	CER8, LACS1, AMP-dependent synthetase and ligase family protein	1.14	2.15	1
At3g16530	Legume lectin family protein	1.14	3.65	1
At2g35680	Phosphotyrosine protein phosphatases superfamily protein	1.15	0	0
At2g41100	ATCAL4, TCH3, Calcium-binding EF hand family protein	1.16	2.67	1
At2g37710	RLK, receptor lectin kinase	1.2	0.95	1
At4g14365	**XBAT34, XB3 ortholog 4 in Arabidopsis thaliana**	1.21	3.5	1
At4g36220	CYP84A1, FAH1, ferulic acid 5-hydroxylase 1	1.22	0	0
At3g50740	UGT72E1, UDP-glucosyl transferase 72E1	1.23	2.38	1
At5g24690	Protein of unknown function (DUF3411)	1.24	0	0
At1g51500	ABCG12, WBC12, ATP-binding cassette, D3, WBC12, ABC-2 type transporter family protein	1.25	0	0
At2g18680	unknown protein: FUNCTIONS IN: molecular_function unknown: INVOLVED IN: biological_proce	1.25	0.76	1
At3g24503	ALDH1A, ALDH12A, REF1, aldehyde dehydrogenase 2C4	1.25	0.73	1
At1g65040	RING/U-box superfamily protein	1.26	0	0
At4g02330	ATPMEPCRB, Plant invertase/pectin methylesterase inhibitor superfamily	1.26	4.94	1
At3g16560	ATCCH1, ATTPC1, PDI2, TPC1, TPC1, two-pore channel 1	1.26	0	0
At5g54710	Ankyrin repeat family protein	1.27	NA	NA
At1g12800	Nucleic acid-binding, OB-fold-like protein	1.28	0	0
At3g18830	ATMES5, ATMTS, PMT5, putative O-methyltransferase 5	1.28	1.92	1
At5g20230	**ATBCB, BCB, BCB, SAG14, blue-copper-binding protein**	1.31	3.62	1
At2g31500	ATR4, CYP83B1, RED1, RNT1, SUR2, cytochrome P450, family 83, subfamily B, polypeptide 1	1.35	0.89	1
At5g18470	Curculin-like (mannose-binding) lectin family protein	1.35	3	1
At1g26380	FAD-binding Berberine family protein	1.37	4.36	1
At3g15356	Legume lectin family protein	1.38	NA	NA
At4g13510	AMT1;1, ATAMT1, ATAMT1;1, ammonium transporter 1;1	1.4	0	0
At1g72060	serine-type endopeptidase inhibitors	1.43	4.38	1
At3g20470	ATGRP-5, ATGRP5, GRP-5, GRP5, glycine-rich protein 5	1.45	NA	NA
At5g11670	ATNADP-ME2, NADP-ME2, NADP-malic enzyme 2	1.46	1.87	1
At5g42020	BIP, BIP2, Heat shock protein 70 (Hsp 70) family protein	1.46	NA	NA
At2g18690	unknown protein: FUNCTIONS IN: molecular_function unknown: INVOLVED IN: biological_proc	1.5	3.27	1
At2g43590	**Chitinase family protein**	1.62	2.05	1
At3g09440	Heat shock protein 70 (Hsp 70) family protein	1.63	0	0
At5g48540	receptor-like protein kinase-related family protein	1.67	1.41	1
At5g02500	AT-HSC70-1, HSC70, HSC70-1, HSP70-1, heat shock cognate protein 70-1	1.68	0	0
At5g55450	Bifunctional inhibitor/lipid-transfer protein/seed storage 2S albumin superfamily protein	1.74	2.78	1
At4g02520	ATGSTF2, ATPM24, ATPM24.1, GST2, GSTF2, glutathione S-transferase PHI 2	1.75	NA	NA
At2g43620	**Chitinase family protein**	1.79	5.61	1
At1g02930	ATGST1, ATGSTF3, ATGSTF6, ERD11, GST1, GSTF6, glutathione S-transferase 6	1.87	NA	NA
At1g11610	CYP71A18, cytochrome P450, family 71, subfamily A, polypeptide 18	1.99	0	0
At1g75040	**PR-5, PR5, pathogenesis-related gene 5**	2.37	3.65	1

	EGM	Common EGM/Skirycz	Common EGM/Skirycz (same sense)
GenesUp	199	70	**68**
GenesDown	22	5	**2**
Not in Array	0	36	**36**
Nb Gene common in arrays	221	185	**185**

TAB. S2. **Differentially expressed genes between arHIF[Col] and arHIF[Cvi] under mannitol treatment from CATMA arrays.**

TAB. S3. **List of primers and vectors used.**

Table S3: List of the primers and vectors used in this analysis:

qPCR			FP	RP
	At1g75040	PR-5	TCACATTCTCTTCCTCGTGTTC	AGGGCAATTGTTCCTTAGAGTG
	At2g43620	chitinase	TGAGATTGCCGCTATGTTGG	AGTCTTTTCCCGGTGTACAAG
	At2g43590	chitinase	ACCCGTGACTCTTTCGTTAAC	GTGTTGCTCCGTTAATCTCTTC
	At5g20230	ATBCB	GTAAAACTTTCCGTGTAGGCG	GTCATGTGGCTAATGGGTTTC
	At2g38870	PR-6 type	TGTCGACCGAATGTCCTAGG	ACTCCCATCCAAAATCACGG
	At2g39200	ATMLO12	GCAGCTACATCACTCTACCTC	ATCCTGAGTGTCCATGTTTCG
	At3g28930	AIG2	TGTATGGAGAATGGGATTTCGAG	CAAACTCTTCCCTACTTCTCCC
	At4g14365	XBAT34	GGTTTTCTGAGGCATGTAAAGG	AAGGAAGTGACATGGCGAC
	At2g38470	ATWRKY33	AGATGGAGAAAATACGGCCAG	TTACTGCTCTCATGTCGTGTG
	At2g25000	ATWRKY60	ACGGGCAAAAGATTACGAGAG	TGTGGTTATGTGTCCCTTCG
	At1g75030	ATLP-3	CGGAATTTTGCTGTACTGGTG	AGATCCGGTACAAGTGAAAGTG
	At2g25490	EBF1	GTCTATTCGTAACTGCCCTGG	GACTCTGTTATCCCCTTGAGC
	At5g47230	ATERF-5	TGTGACTGGGATTTAACGGG	CAACTGGGAATAACCAAACGG
	At5g25350	EBF2	TTGGGAATGAGGGTTTAAGGG	GCATCTGGAGTTTCACTTTCG
	At1g14920	GAI	ATGTCGGAGGTTTACTTGGG	TTCATGACGCTCAACTCGG
	At1g13320	PP2A	TTTGTGAAGCTGTAGGACCG	CGAGTTCAGGGTTTAAAATGCG
	At1g13440	GAPDH	TTGGTGACAACAGGTCAAGC	AAACTTGTCGCTCAATGCAATC

Cloning		Use	Attb1 + + Specific sequence	Attb2 + + Specific sequence	Destination vectors	From	Bacteria/Plant antibiotic selection	Transformed Genotype
	EGM2 ExtraCellularDomain	-	GGGGACAAGTTTGTACAAAAAAGCAGGCTTCGAAAGAGCTTTATTTTC	GGGGACCACTTTGTACAAGAAAGCTGGGTTCTATTCTGAATGT	pDest17	InVitrogen	Carbenicilline	-
	EGM1 (+stop)	Transgenic complementation	GGGGACAAGTTTGTACAAAAAAGCAGGCTATGAGACTTCATGAGAGC	GGGGACCACTTTGTACAAGAAAGCTGGGTTTAACGTCCTGAGA	pmDC32	Curtis2003PlantPhysiol	Hygromycin	arHIF[Col]; arHIF[Cvi]
	EGM2 (+stop)	Transgenic complementation	GGGGACAAGTTTGTACAAAAAAGCAGGCTATGAGACTTCATGAGAGC	GGGGACCACTTTGTACAAGAAAGCTGGGTCTAACGTCCTGTAA	pmDC32	Curtis2003PlantPhysiol	Hygromycin	arHIF[Col]; arHIF[Cvi]
	pEGM1	GUS	GGGGACAAGTTTGTACAAAAAAGCAGGCTGCATACATATTCTCAAGCC	GGGGACCACTTTGTACAAGAAAGCTGGGTCATGAAGTCTCATA	pBi101-R1R2	F. Divol. J.-C. Palauqui and B. Dubreuc	Kanamycin	Col-0
	pEGM2	GUS	GGGGACAAGTTTGTACAAAAAAGCAGGCTGGGTCGTTGGAACATAGT	GGGGACCACTTTGTACAAGAAAGCTGGGTCTCAAGCTCTCATC	pBi101-R1R2	F. Divol. J.-C. Palauqui and B. Dubreuc	Kanamycin	Col-0
	pEGM1:EGM1 (+stop)	Transgenic complementation	GGGGACAAGTTTGTACAAAAAAGCAGGCTGCATACATATTCTCAAGCC	GGGGACCACTTTGTACAAGAAAGCTGGGTTTAACGTCCTGAGA	pgwB1	Nakagawa2007JBioscBioeng	Kanamycin / Hygromycin	Salk_058300; WiscDsLox426E06
	pEGM2:EGM2 ExtraCellularDomain	-	GGGGACAAGTTTGTACAAAAAAGCAGGCTGGGTCGTTGGAACATAGT	GGGGACCACTTTGTACAAGAAAGCTGGGTTCTATTCTGAATGT	-	-	-	-
	pEGM2:EGM2 (+stop)	Transgenic complementation	-	-	pgwB1	Nakagawa2007JBioscBioeng	Kanamycin / Hygromycin	WiscDsLox426E06
	pEGM1 to EGM2 (8kb)	Transgenic complementation	-	-	pgwB1	Nakagawa2007JBioscBioeng	Kanamycin / Hygromycin	WiscDsLox426E06
	EGM_amiRNA_F (I)	amiRNA line	gaTCATATCACAGTTTGTACGGGtctctcttttgtattcc					
	EGM_amiRNA_R (II)	amiRNA line	gaCCCGTACAAACTGTGATATGAtcaaagagaatcaatga		pGREENII0000			Col-0
	EGM_amiRNA*_F (III)	amiRNA line	gaCCAGTACAAACTGAGATATGTtcacaggtcgtgatatg				Kanamycin	
	EGM_amiRNA*_R (IV)	amiRNA line	gaACATATCTCAGTTTGTACTGGtctacatatatattcct					

Genotyping		Targeted gene	WT band		TDNA band		Enzyme (CAPS)
			LP	RP	LB	RP (if different from the one for WT band)	
	SALK_206891	At1g11280	ATCACAAAAATGGAAAGGCTG	TGTTTGGTCAACCAGAAGACC	ATTTTGCCGATTTCGGAAC		
	WiscDsLoxHs015_10B	At1g11280	GGCTGGGTTCTGGACTTTATC	GGGCAAAGAAAACCAATTCTC	AACGTCCGCAATGTGTTATTAAGTTGTC		
	SALK_122320	At1g11290	TCCTTCGTCCAACATCTTTTG	GCAGAACAAAAGCATGAGCTC	ATTTTGCCGATTTCGGAAC		
	SALK_008433	At1g11300 (promoter)	CTCAGCGACAGTAGAATTGGC	CTCGGAGAGAAAGTGACGATG	ATTTTGCCGATTTCGGAAC		
	SAIL_150_H02	At1g11300	GGGGATTAGGGATTTGAGGA	CTTCTCCTGCAATGCAACAA	GCTTCCTATTATATCTTCCCAAATTACCAATAC/	TGCTCTGCATCTCTTCCTTTC	
	SALK_058300	At1g11300	AAGAAGTTTTGCCCTCCTTGAG	GATCGGTACAATATTGGCAGGC	ATTTTGCCGATTTCGGAAC		
	SALK_044069	At1g11300	GCCTTAGCTGGTGGTAATAAAGG	GCATATGACATGGTGGCTGA	ATTTTGCCGATTTCGGAAC	GATTGCTGTGAAGAGGCTGTC	
	WiscDsLox426E06	At1g11305	AAGAAGTTTTGCTTCCCTTGG	CTGACTTTGCAAGAAGATCCG	AACGTCCGCAATGTGTTATTAAGTTGTC		
	SALK_050191	At1g11310	TCAGGACGTTAAAACCTGCAG	TGATGACCTCTTCTGGTTTGG	ATTTTGCCGATTTCGGAAC		
	SALK_079850	At1g11310	GAAGTCGTGGGAGGAAGAGAC	TGCAAGTTGAAATGCATCTTG	ATTTTGCCGATTTCGGAAC		
	SALK_143489	At1g11330	TACTCCGGTGACATATAGCCG	TGCAATCATTTTTGGGAAGTG	ATTTTGCCGATTTCGGAAC		
	SALK_099776	At1g11350	ATAACCGCTGTGGTGTTATGC	TATGCAGGAGTTTTCTGGCAC	ATTTTGCCGATTTCGGAAC		
	SALK_076477	At1g61610	TCCAAGCTTGTTCTCTTCAGC	CTTGAGGTTTTGGATTAGGCC	ATTTTGCCGATTTCGGAAC		
	SAIL_1212_G06	At3g16030	GCAAGAGCTCGTTTCAGCTT	CCTTTTCGTTCCCTTGGTCT	GCTTCCTATTATATCTTCCCAAATTACCAATAC/	TTACCTCTTTGTAACGTGCGG	
	SALK_067606	At4g27290	ACCTTGTCCAAGTTTGTTCCC	TGTTGTCTTCAGGACTGGTCC	ATTTTGCCGATTTCGGAAC		
	SAIL_629_C11	At4g27290	TGAGAAATACCCGTCGATTTG	TTTTTGTTCCGTTTCGATCAC	GCTTCCTATTATATCTTCCCAAATTACCAATACA		
	SALK_145131	At4g27300	CCATGAGATTCCAACCAACTG	CCCAAAAGATTTAGCCAATCC	ATTTTGCCGATTTCGGAAC		
	N8052	At3g20770	GAGCAAGCTAGGAGGAAGAAATGTCTAG	TTTAGGCAAACCAAGTTGGATGCCAC	-		BsuRI
	SALK_020997	At2g25490	CTGTGGCATTCTTCAAGC	TCTTGTCAGGGATCAAGG	ATTTTGCCGATTTCGGAAC		
	GK-067H03	At5g47230	ATGGCGACTCCTAACGAAGTA	AACAACGGTCAACTGGGAATA	CCCATTTGGACGTGAATGTAGACAC		
	CAPS marker (polymorphism e)	At1g11305	GACTCCTCTGGAGTGATTTCAA	GCATATGACATGGTGGCTGA			AluI
	CAPS marker (polymorphism f)	At1g11305	AAAGGGTTTAGGCCAAGGAA	ATTCTGGTTCGCTGGCTTC			HphI
	MSAT103527	-	CCATACCATGAGAGGCGTTA	GCAACATCTGTGTTCGTCGT			
	IND103875	-	ATTGGTCAGGAAGTGGCAAG	TGTCTCTGGTCGCTACATGG			
	MSAT103897	-	GGCCAAACGGAGTTGACATA	AGTTACCACCACTATCACTCTTATCA			
	MSAT103776	-	AAAACTAATGTGGTGATGAATGG	CATGATTTTGAGCCACTTGGT			

Table S4: List of the genetic material used in this analysis.

POPULATIONS			Av number (parent)				
	Cross	AV number (pop)	femelle	male	From	Type	MSAT used for genotyping *EGM*
RILs and F2	Cvi-0 x Col-0	8RV			VASC	RILs	
	Sha x Col-0	13RV			VASC	F2	MSAT103527
	Col-0 x Sha	127RV			VASC	F2	MSAT103527
	Sha x Cvi-0	39RV			VASC	F2	IND103875
	Cvi-0 x Sha	70RV			VASC	F2	IND103875
	Cvi-0 x Bur-0	65RV			VASC	F2	MSAT103897
	Col-0 x Bur-0	20RV			VASC	F2	MSAT103527
	Cvi-0 x Ct-1	66RV			VASC	F2	MSAT103527
	Col-0 x Ct-1	7RV			VASC	F2	MSAT103527
	Blh-1 x Cvi-0	87RV			VASC	F2	IND103875
	Cvi-0 x Bllh-1	64RV			VASC	F2	IND103875
	Blh-1 x Col-0	21RV			VASC	F2	IND103875
	Cha-1 x Col-0	-	227AV	186AV	In this study	F2	MSAT103776
	Col-0 x Cha-1	-	186AV	227AV	In this study	F2	MSAT103776
	Cha-1 x Cvi-0	-	227AV	166AV	In this study	F2	IND103875
	Cvi-0 x Cha-1	-	166AV	227AV	In this study	F2	IND103875
	Ze-0 x Col-0	-	529AV	186AV	In this study	F2	MSAT103897
	Col-0 x Ze-0	-	186AV	529AV	In this study	F2	MSAT103897
	Ze-0 x Cvi-0	-	529AV	166AV	In this study	F2	IND103875
	Cvi-0 x Ze-0	-	166AV	529AV	In this study	F2	IND103875
	Tou-1 x Col-0	-	646AV	186AV	In this study	F2	IND103875
	Col-0 x Tou-1	-	186AV	646AV	In this study	F2	IND103875
	Tou-1 x Cvi-0	-	646AV	166AV	In this study	F2	MSAT103897
	Cvi-0 x Tou-1	-	166AV	646AV	In this study	F2	MSAT103897
MUTANTS							
	Line	**locus**	**location**	**name**	**Background**	**From**	**Mutation type**
EGM region:							
	SALK_206891	At1g11280	Coding region		Col-0	NASC	tdna
	WiscDsLoxHs015_10B	At1g11280	Coding region		Col-0	NASC	tdna
	SALK_122320	At1g11290	Coding region		Col-0	NASC	tdna
	SALK_008433	At1g11290	promoter		Col-0	NASC	tdna
	SAIL_150_H02	At1g11300	Coding region		Col-0	NASC	tdna
	SALK_058300	At1g11300	Coding region	*(egm1)*	Col-0	NASC	tdna
	SALK_044069	At1g11300	Coding region		Col-0	NASC	tdna
	WiscDsLox426E06	At1g11305	Coding region	*(egm2)*	Col-0	NASC	tdna
	SALK_050191	At1g11310	Coding region		Col-0	NASC	tdna
	SALK_079850	At1g11310	Coding region		Col-0	NASC	tdna
SD1 family							
	SALK_143489	At1g11330	Coding region		Col-0	NASC	tdna
	SALK_099776	At1g11350	Coding region		Col-0	NASC	tdna
	SALK_076477	At1g61610	Coding region		Col-0	NASC	tdna
	SAIL_1212_G06	At3g16030	Coding region		Col-0	NASC	tdna
	SALK_067606	At4g27290	Coding region		Col-0	NASC	tdna
	SAIL_629_C11	At4g27290	Coding region		Col-0	NASC	tdna
	SALK_145131	At4g27300	Coding region		Col-0	NASC	tdna
Hormone							
Ethylène	N8052	At3g20770	Coding region	ein3-1	Col-0	NASC	ems
Ethylène	SALK_020997	At2g25490	Coding region	ebf1-3	Col-0	NASC	tdna
Ethylène	GK-067H03	At5g47230	Coding region	erf5	Col-0	NASC	tdna

TAB. S4. **List of genetic materials used.**

10. GENERAL DISCUSSION AND PERSPECTIVES

10.1 EGM1 and EGM2 and biotic stress

10.1.1 Innate immunity in plants

Plants are constantly exposed to a vast diversity of microbes–potentially pathogenic–and so evolved a series of intricate mechanisms to prevent or reduce pathogen infections (i.e. resistance mechanisms) or attenuate their fitness consequences (i.e. tolerance mechanisms) [285]. Whereas little is known about the molecular mechanisms mediating tolerance, several mechanisms responsible for resistance are known. First, resistance mechanisms consist in multiple physical and chemical barriers such as cuticles, waxes, trichomes, cell walls and antimicrobial compounds that limit pathogen growth at the surface of plant cells [285]. Then, if pathogens circumvent these barriers, inducible responses are triggered by the recognition of generally conserved microbe-associated molecular patterns (MAMPs) [22] by extracellular surface receptors named PRR (Pathogen Recognition Receptors) such as FLS2 (FLAGELLIN-SENSING 2) that recognises flagellin or EFR that recognises the elongation factor EF-Tu (ELONGATION FACTOR [285, 286]. Similar danger signals may also arise from damage-associated molecular patterns (DAMPs) such as cell wall and waxes derived molecules that come from the damage of plant molecular structures. The signalisation triggered by PRR leads to the activation of basal defences (also called PTI (or MTI) for PAMP- (MAMP-) triggered immunity) that are relatively weak and transient but sufficient to contain most microbes [285, 286]. Nevertheless some pathogens are able to suppress basal immunity by secreting virulence factors (also known as effectors) that prevent MAMP perception and signalling. To counteract this pathogen response and contain pathogens spread throughout the plant, plants proteins encoded by R-genes can recognise specifically pathogen effectors and trigger the hypersensitive response that consists in a rapid apoptotic cell death that can result in necrosis [285, 286]. The resistance acquired via effector–triggered immunity (ETI) is known to be more specific, more prolonged and more robust than the one triggered by PTI [287]. However these two pathways share several components [287]. This includes similar early responses and signalling machinery such as transcriptional changes, reactive oxygen species (ROS) production, MAP kinase signalling ac-

22. The term of PAMPs for pathogen-associated motifs can also be used but the term of MAMPs is preferred as those conseved molecular patterns are also shared with non pathogenic microbes.

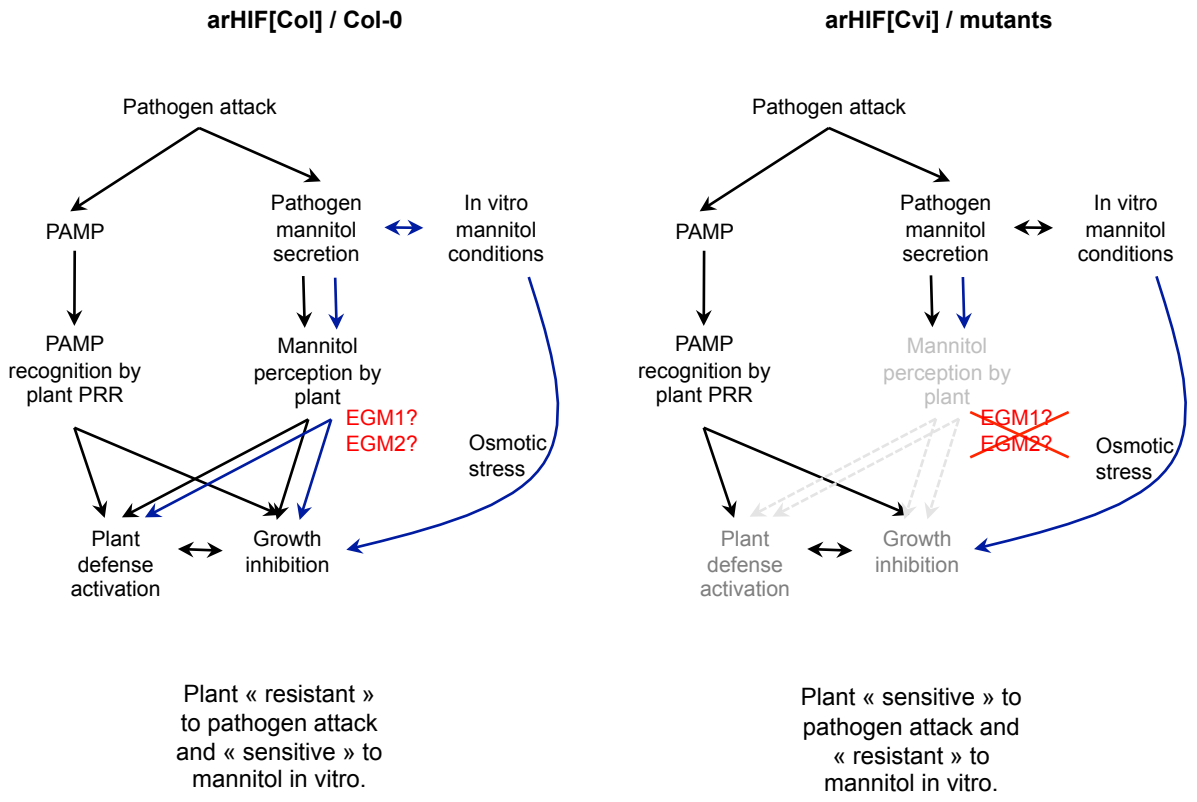


Fig. 32. EGM1 and EGM2 model of action regarding *in vitro* mannitol stress and biotic stress. In Col-0 and in the arHIF[Col], during biotic attack (black arrows), the mannitol produced and secreted by the pathogen could be perceived by the plant that would induce via EGM1and EGM2 a response contributing to the growth inhibition and plant defense activation associated to MAMP triggered immunity. When mannitol is applied *in vitro* (blue arrows), it somehow simulates mannitol pathogen secretion, which activates EGM1 and EGM2 pathways resulting in partial induction of defense responses and growth inhibition in addition to the effect induced by osmotic stress. Mutation of EGM1 or EGM2 results in the absence of mannitol perception pathway so that plant are relatively more 'sensitive' to pathogen attacks but appear more 'resistant' to mannitol stress *in vitro*.

tivation and hormonal signalling via salicylic acid (SA) and jasmonic acid (JA) / ethylene (Et) cross talks as well as similar downstream responses such as cell wall fortification, production of antimicrobial secondary metabolites and accumulation of PR-proteins. As so the distinction between the two is not always straightforward. Finally, the local perception of MAMPs can also induce systemic defence responses to protect undamaged tissues against subsequent invasion by the pathogens [286].

10.1.2 Mannitol signalling pathway in plant immunity

From our analysis of *EGM* QTL, we suggest the following model. During infection, the mannitol produced and secreted by pathogens could be sensed directly or indirectly by plants which in turn could activate a pathway that synergistically contributes to the plant defence

activation and growth repression shared with other pathways of plant innate immunity (i.e. the basal PAMP-triggered and/or the effector triggered immunity). Under *in vitro* mannitol stress, this mannitol-responsive pathway is constitutively activated and also leads to growth repression. However, when the EGM1 and/or EGM2 proteins –that would contribute to mannitol signalling– are mutated, the plants can't respond anymore to mannitol so that those mutants are more sensitive to pathogens' attacks but grow better on mannitol *in vitro* (figure 32).

From our results and the known responses generally triggered by PTI and ETI, we don't know if mannitol is perceived directly or indirectly by plants and if this polyol can be considered as a MAMP/effector or not. Under *in vitro* condition, mannitol is responsible for the induction of stress-responsive genes (Publication Fig. 4, Fig. S6, Table S2, [288, 289]) that likely turn on shoot growth inhibition. However no evidence of necrosis or callose accumulation have been observed (data not shown). Besides, the growth response under mannitol treatment is observed from relatively high concentration (5-10mM, Publication Figure S5) compared to standard MAMPs such as flg22 (from 0.01-10 μM). The timing of mannitol response is also unknown. It would have been interesting to test for other classical PTI/ETI defence responses such as ROS production, medium alkalinization,camalexin and hormones accumulation...

SA, JA and Et are the classical hormones involved in biotic stress response, SA being more specifically involved in the defence against biotrophic pathogens[23] and JA/Et against necrotrophic pathogens[24][290]. We tested the role of the JA, SA and Et hormonal pathways in shoot growth repression under mannitol stress by crossing *egm1* and *egm2* mutants to several lines mutated for key proteins within those pathways. We also tested abscisic-acid (ABA) hormonal pathway that is commonly associated with abiotic stress tolerance but which role in biotic stress tolerance is becoming increasingly evident [291]. No consistent interaction between EGM and hormonal pathways has been observed under mannitol stress (figure 33) suggesting that the shoot growth repression due to EGM signalling under mannitol stress is not completely dependent from those hormonal pathways. Those results could also be explained by redundancy. Indeed, a recent paper show that *ERF5* and *ERF6* genes are likely redundant and that the double *erf5* x *erf6* mutant shows an enhanced shoot growth phenotype under mannitol stress similar to the one of *egm* mutants [292]. It would be interesting to cross this double mutant with *egm* mutants to check if their are involved in the same pathway or not. Another possibility could be that JA, SA and Et hormonal pathways are able to compensate each others (and hence we would not observe direct control of one of them in isolation from the others), as observed after infection with Pseudomonas syringae and Alternaria brassicicola [293]. The effect of auxin and gibberellin hormones would need to be further analysed as well.

23. Biotrophic pathogens use nutrients derived from the host cells without killing them.
24. Necrotrophic pathogens use nutrients derived from plants dead tissues.

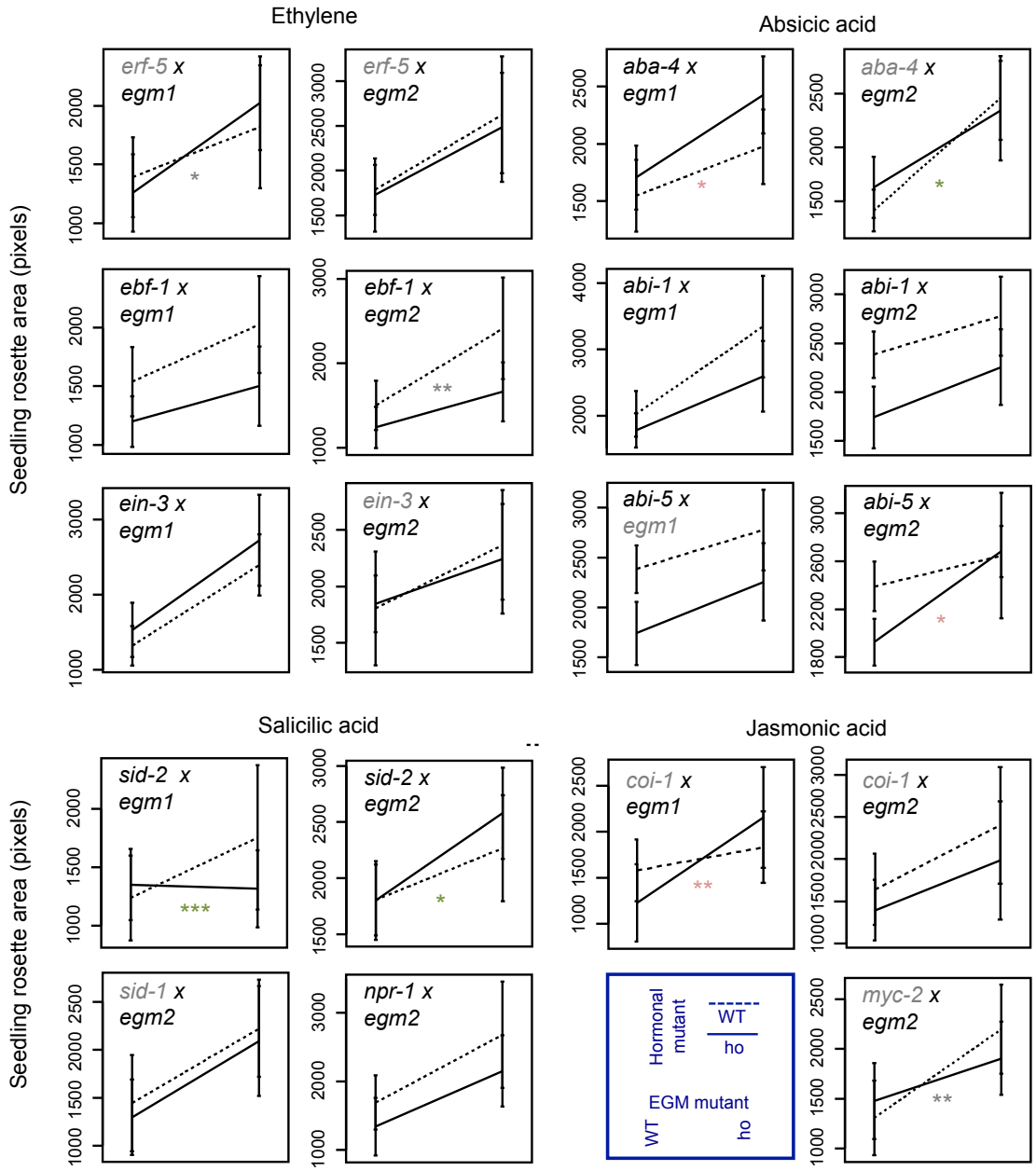


Fig. 33. Effects of ethylene, abscisic acid, salicilic acid and jasmonic acid on EGM phenotype. *egm1* and *egm2* mutants have been crossed with several mutants of the biosynthetic and/or signalling hormonal pathways. Interaction plots of the WT, *EGM* ho mutant, hormonal ho mutant and double mutant for each cross are represented. For each plot, the absence of effect of *egm* mutation or hormonal mutation is indicated in grey in the name of the cross. The significance of the effect of the interaction between *egm* and hormonal mutations is indicated by stars (ANOVA; $*p < 0.05$, $**p < 0.01$, $***p < 0.001$). Grey stars are interactions that were not confirmed in a second experiment. Pink stars are the ones that are unlikely because the effect of *egm* mutation is not clear in the WT background of the hormonal mutant. Green stars are the interactions that would need to be confirmed in a second experiment. For the crosses with *abi-1* (Ler background) and *abi-5* (Ws-0 background) multiple lines have been phenotyped because of putative QTLs segregating in the F2 Ler x Col-0 and Ws-0 x Col-0 respectively. For other crosses, only one line have been phenotyped. Error bars represent the standard deviation observed in at least 30 plants.

Finally, if mannitol is perceived directly as a signal and induce change in stress responsive genes that turn on growth inhibition, we can wonder why no shoot phenotype has been observed in most of the plants transformed with either bacterial mannitol-1P-deshydrogenase (*mt1d*) or celery mannose-6-phosphate reductase (*M6PR*) (including *A. thaliana*) under control conditions [289, 294, 295, 296, 297, 298, 299]. One reasonable explanation could be that those transgenic plants, do not accumulated enough mannitol to trigger the growth inhibition phenotype (although part of the signaling response could be activated). This hypothesis is consistent with the observation that in wheat the growth defects observed in *mt1d* transgenics are more severe in plants accumulating more mannitol although this phenomenon could also be due to perturbation of sugar metabolism [300]. In our case, we observed an increased in sugar content in the arHIFs grown on mannitol (sacharose, sucrose, fructose and mannose, data not shown) but we did not detect any interaction with EGM functionality suggesting that the growth phenotype observed under mannitol stress is independent from this sugar perturbation.

10.1.3 Which pathogens and in which plants?

Mannitol is a carbohydrate commonly occurring in pathogenic and non-pathogenic fungi and bacteria but also in higher plants [284, 301] where it serves as carbon storage and translocation form, as well as osmoticum. As a result, we can wonder if all plants can sense mannitol and respond to it through the activation of plant defences.

First, we can think that only non-mannitol producing plants such as Arabidopsis or tobacco developed mannitol perception mechanism. In Arabidopsis, the fact that *EGM2* is not retrieved in *A. lyrata* and *A. halleri* suggests that either the two latter species do not produce mannitol or that the mannitol responsive pathway we are describing is specific of *A. thaliana* or that in those species a single EGM receptor might be sufficient to trigger mannitol-signalling response. Tobacco may have also acquired a mannitol-sensing pathway, which leads, through the activation of the SA signalling, to the release of a mannitol deshydrogenase in the extracellular space where the mannitol released by the pathogen is likely to quench the ROS produced by the plant during infection [302]. However, it is also possible that even mannitol-producing plants perceived the mannitol either constitutively or through concentration thresholds for example. Interestingly the celery mannitol deshydrogenase (converting mannitol to mannose) is upregulated by SA [303, 304]. However it remains to be determined if this upregulation is just a remobilisation of carbohydrate storage or really makes sense to limit the growth of pathogens. Finally, as mentioned earlier, it is likely that mannitol pathway acts synergistically with the activation of other PRR so that its sole presence in most plants won't trigger an important defence response.

The importance of mannitol production and secretion upon infection might also vary depending on the pathogen. As noted already, one of the often-proposed reasons for pathogen mannitol secretion is the quenching of ROS species produced by plants at the infection site [305, 306, 307, 308]. However, because necrotrophic pathogens could take advantage of host cell death, they may have no obvious interest to quench ROS [309]. Nevertheless mannitol is required for normal disease development by *Alternaria alternata*, a necrotrophic pathogen, on tobacco [308]. In Botrytis, we show that high level of mannitol and the absence of EGM signalling pathway in the host results in a higher growth of the pathogen. Overall, we don't know if the primary role of mannitol production in necrophytic pathogen is carbohydrate storage [310, 311, 312] or ROS quenching.

Because Botrytis does not affect *A. thaliana* in natural conditions and to confirm the role of mannitol signalling pathway in pathogen defence, we decided to test another necrotrophic pathogen, *Sclerotinia sclerotiorum*, known to infect *A. thaliana* in natural environmental conditions [313] and to produce mannitol [310, 314]. This pathogen was particularly interesting as from aranet database *EGM2* and maybe *EGM1* are expected to be highly expressed in petals, which in canola are an important vector of plant infection by *S. sclerotiorum*. Dominque Robby and Claudine Balagué from the laboratory LIPM (INRA Toulouse) performed three experiments for us where they tested the sensitivity to *S. sclerotiorum* of different genotypes 28 days after sowing [315]. The two first experiments were performed in greenhouse conditions and no consistent difference in sensitivity was observed between the mutants, WT and overexpressors. Overall, too few plants were used (3-5) due to space limitation and environmental conditions were not ideal (very cold and very hot periods respectively) so a third experiment in phytotron with more plants (from 9 to 14) was performed. No difference between the overexpressors and WT (arHIF[Cvi]) was observed, but those genotypes were overall much more sensitive than the one in a Col-0 background likely as a result of the (partial) Cvi-0 genetic background of the arHIF (figure 34A). Indeed Cvi-0 is highly sensitive to *S. sclerotiorum* [315]. In this experiment, a significant difference was observed between the *egm1* and *egm2* mutants and WT plants, the mutants being more sensitive to Sclerotinia, which is in accordance with the results observed with Botrytis (figure 34A).

Then we analysed the level of expression of *EGM1* and *EGM2* in infected Col-0 plants. *EGM1* was upregulated during infection as expected from our mannitol perception model. However *EGM2* was downregulated relatively early after inoculation (figure 34B). This result was somehow surprising as it was the first time that we saw the two genes behaving differently, suggesting that within the 250bp promoter of *EGM2* a specific regulatory element allows him to have a specific transcriptomic behaviour.

Finally, our model suggested that both receptors act in the same pathway and so from this model we expected that both genes would respond similarly to the mannitol produced by

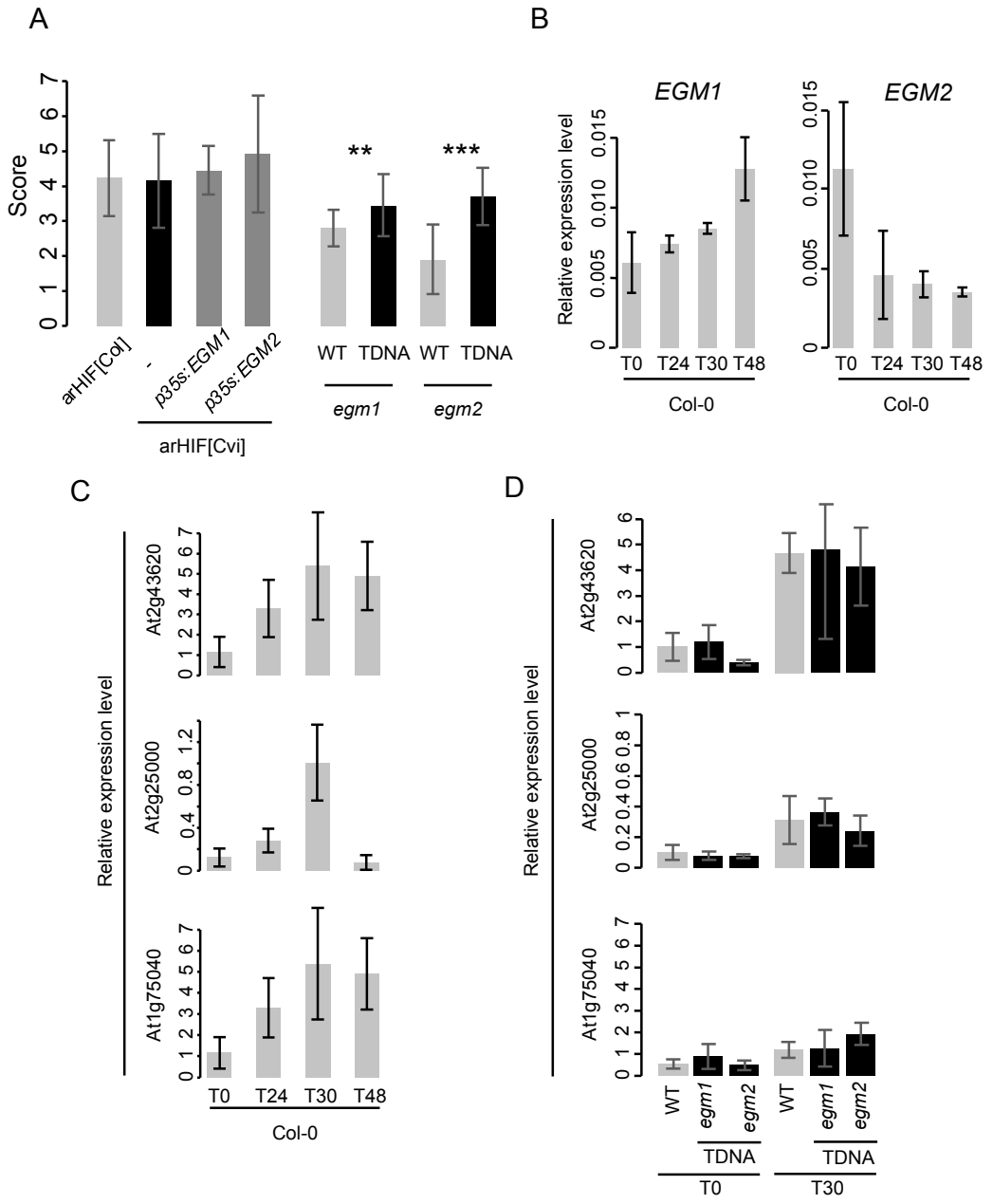


Fig. 34. *EGM1* and *EGM2* response to *Sclerotinia sclerotiorum*. A. Disease score of 4 weeks old plants, 5 days after inoculation of one leaf with *S. sclerotiorum*, as measured by *Perchepied et al. 2010* [315]. Error bars correspond to the standard deviation observed in 9 to 14 plants grown in different trays. For *egm1*, data from two independent T-DNA lines were pooled. Stars represent significant differences observed from a Kruskal Wallis test. Importantly, a second biological replicate gave no significant results. B-C. Relative expression level of *EGM1* and *EGM2* (B) and 3 mannitol responsive genes (C) in 4 weeks old Col-0 plants, before and 24h, 30h and 48h after inoculation of three leaves with *S. sclerotiorum*. D. Comparison of the relative expression level of 3 mannitol responsive genes in 4 weeks old plants with functional or nonfunctional EGM1 and/or EGM2 receptors, before and 30h after inoculation of three leaves with *S. sclerotiorum*. *At2g43620*: chitinase, *At1g75040*: PR-5, *At2g25000*: WRK60. *GAPDH* and *PP2A* (B) or *UBC21* and *At2g28390* (C-D) endogenous controls were used to determine the expression levels of the different genes using the $2^{(Ct_{gene} - mean(Ct_{EndogenousControls}))}$ formula. Error bars represent the standard deviation observed in three pools of three leaves sample from three independent individuals during the same experiment.

pathogens. Those results suggest that some elements are missing in our model and that the relations between *EGM* genes may be more complicated than suggested in the paper. The two EGM receptors are likely to share common functions but also may also be involved in independent pathways. Conversely, it is possible that the regulation observed here is fortuitous, specific to this experiment or specific to *S. sclerotiorum* pathosystem. To try and further link *in vitro* mannitol stress, biotic stress and EGM1/EGM2 functionality, we wanted to see if the genes that were upregulated under mannitol stress in WT background but not in *egm1* and *egm2* mutants behave similarly after inoculation with *S. sclerotiorum*. We confirm that *PR5*, *WRK60* and *At2g43620* gene encoding a chitinase, were upregulated under biotic stress (figure 34C) but this induction was independent from EGM1 and EGM2 functionality (figure 34D). It is very likely that these three genes are (also) activated via other pathways in response to pathogens.

Overall, the role of mannitol for the pathogen and the interest of a mannitol-sensing pathway in plant hosts might be specific to each pathosystems and requires more analyses.

10.2 What are the roles of EGM receptor-like-kinases within mannitol signalling pathway?

10.2.1 The role of plant receptor like kinases in plant innate immunity

The RLK/pelle protein family is a large family of kinases characterized by a dramatic expansion in plants compared to animals [316]. In *A. thaliana*, over 600 members with two possible configurations have been isolated. Receptor-like cytoplasmic kinases (RLKC) are characterised by a transmembrane and/or Ser/Thr kinase domains whereas 75% of the other members (RLKs) are characterised by an intracellular Ser/Thr kinase domain, a transmembrane domain and an extracellular domain (ECD) [317]. The ECD is composed of various motifs implicated in various functions such as protein-protein or protein-carbohydrate interactions. Thus the ECD is likely essential for the perception of various signals produced by the plant or derived from microbes. RLKs are involved in a wide range of developmental processes in plants as well as in abiotic and biotic stress responses. Among the best-characterised RLKs are the brassinosteroid receptor BRI1 (BRASSINOSTEROID INSENSITIVE 1) important for growth regulation, the receptor of CLAVATA3, CLAVATA1 (CLV1), essential for meristem size regulation and the flagellin receptor FLS2 involved in immune responses. But more and more RLKs are being discovered as PRR (or MAMPs receptors) and the expansion of the RLK/Pelle family is often perceived as a response to highly variable and rapidly evolving biotic agents [318, 319, 320]. Thus RLKs are particularly important for plant innate immunity.

The flagellin/FLS2 ligand-receptor pair can be used as an example to illustrate the general

molecular events associated with the activation of RLKs. In Arabidopsis, a 22-amino acid peptide (flg22) corresponding to a highly conserved motif of the flagellin N-terminal region is sufficient to induce plant defences [321] and was used for the identification of FLS2 in a growth defect-suppressing mutant screen [322, 323]. FLS2 has an extracellular domain composed of 28 LRR (for Leucine Rich Repeat) motifs organised in beta-strand/beta-turns structures with LRR 9 and 15 particularly important for flg22 and flagellin binding to FLS2 [324]. The interaction of flg22 with FLS2 is accompanied with a hetero-dimerization with BAK1 (BRI1-ASSOCIATED-KINASE 1), another receptor like kinase originally identified as a co-receptor of BRI1. The transphosphorylation of their kinase domain [325] lead to the phosphorylation of the kinase domain of a third partner, BIK1 (BOTRYTIS-INDUCED-KINASE 1), which once activated is released from the complex and activates downstream targets [326]. The activation of FLS2 is followed by its internalisation in the endomembranous system, which could lead to a new signalling function and/or the degradation and recycling of the receptor [286, 327, 328]. Finally, FLS2 is degraded via the proteasome and avoids a too long or too strong immune response [329].

Overall, the activation of RLKs is often performed in complexes and involved the activation of several partners through auto- and trans-phosphorylations of kinase domains. After activation and transduction of the signal, activated receptors are internalised and degraded via the proteasome [330].

10.2.2 What is the function of EGMs RLKs?

In this paper, we identified two receptor-like kinases likely contributing to mannitol signalling and to the growth repression observed under mannitol stress. Some of the important questions we did not answer are whether mannitol is directly or indirectly perceived by the plant and where exactly in the pathway are the two receptors acting? Indeed these receptors could be more or less directly involved in the perception of mannitol or be the receptors of a downstream signal.

To test the possibility that EGM1 and EGM2 were mannitol receptors we wanted to know whether or not those receptors could be internalised in the endomembranous system under mannitol treatment. First we wanted to know whether the RLKs were localised at the plasma membrane or not and if the Cvi allele of EGM2 was mislocalised or not. To do so, we fused the green fluorescent protein (GFP) to the C-terminal end of $EGM2^{[Col]}$ and $EGM2^{[Cvi]}$ alleles and expressed these constructs in the arHIF[Col] and in the arHIF[Cvi] under the CaMV-35s promoter (figure 35A). We did not perform N-terminal fusions because a signal peptide was predicted in that region in the two RLKs. Unfortunately, *p35s:EGM2[Col]-GFP* construct did not complement the phenotype of the arHIF[Cvi] under mannitol stress (figure 35B) although a GFP signal could be observed (figure 35C). Overall, the GFP signal observed in several independent insertion lines was weak suggesting that the transgenes were not highly expressed.

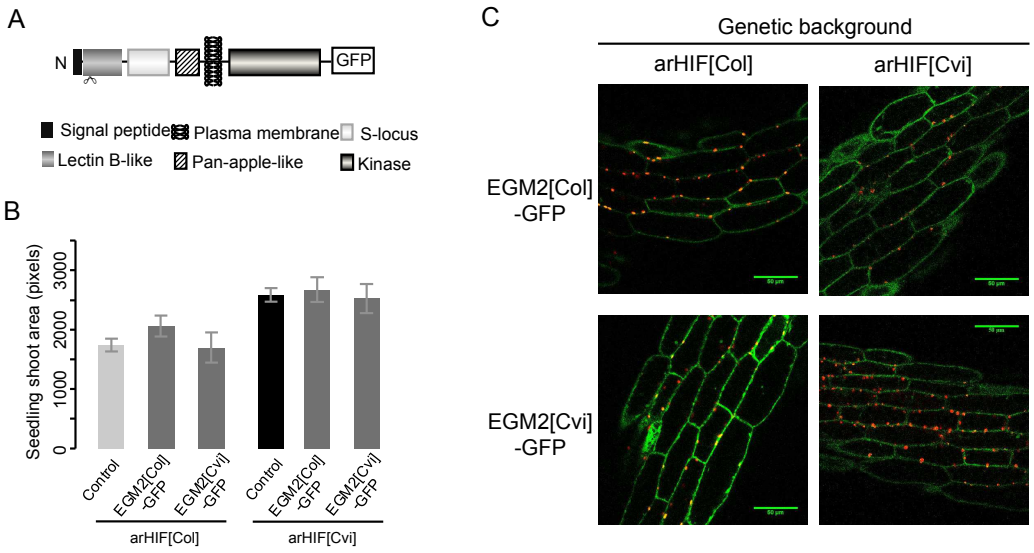


Fig. 35. Analysis of EGM2 localisation using GFP fusions. A. Because of the presence of a N-terminal signal peptide, GFP was fused to EGM2 C-terminal end only and expressed in the arHIF[Col] and arHIF[Cvi] under the pCaMV-35s promoter. B. Shoot area of 12 DAS seedlings of the arHIF[Col] and arHIF[Cvi] transformed with *p35s:EGM2[Col]-GFP* . *p35s:EGM2[Col]-GFP* fusion do not complement the arHIF[Cvi] suggesting that the constructs are not functional. Error bars represent the standard deviation observed in different individuals of the insertion line with the brightest GFP signal (per construct). C. GFP signal observed in the hypocotyles of the insertion lines mentioned in B. under Leica SP2 confocal microscope. Red signal corresponds to autofluorescence.

We should have tried to express EGM1-GFP fusion because *EGM1* seems to be more easily overexpressed *in planta* (Publication – Fig. 2B, Fig. 3A). But at that time, we were mainly interested in EGM2 because it was the RLK likely responsible for *EGM QTL* and so we did not perform *EGM1* constructs and transformations. However, because a complementation could be observed with *p35s:EGM2[Col]* construct (Publication – Fig. 2B) we were afraid that the absence of complementation of the arHIF[Cvi] with the *p35s:EGM2[Col]-GFP* construct was due to the GFP preventing either the phosphorylation of the kinase domain or the interaction with other proteins leading to an unfunctional allele. Consequently, we decided to produce an antibody against EGM2 extracellular domain. I produced this domain (without the signal peptide; 1.206kb) fused to a 6xHis tag (*pDEST*^{TM}17) in Origami 2 (Novagen) *E. coli* cells. The peptide was highly insoluble and so was just isolated from membranes by several washings (figure 36A)[25]. The two rabbit polyclonal antibodies (*Totoro* and *Mei*) raised against the recombinant protein (Eurogentec) recognised 50 nmol of the recombinant EGM2 extracellular domain (figure 36B) by western blot. A western blot performed on plant extracts (grown on mannitol) using *Totoro* antibody revealed a signal of the expected size (figure 36C) but no difference in the intensity of the signal was observed between the arHIF[Cvi] and the overexpressors despite similar protein loading. Besides the observed signal was specific of the soluble fraction (S) and not associated with the membranous fraction (Mb) (figure 36C). This could be due to the fact that plants were grown on mannitol and so that the EGM1 and EGM2 were partially degraded in those conditions or that the observed signal was aspecific. Indeed, EGM1 and EGM2 belong to a multigenic family and so our polyclonal antibody was likely to target the other members of the family (or other proteins characterised by the same protein domains).

Still, we decided to test this antibody in immunolocalisation experiments[26] because we were thinking that a specific signal could be observed in some tissues in the overexpressing lines compared to the WT. In all the lines tested, a strong signal, likely membranous, was observed at the root tip using *Totoro* rabbit polyclonal antibody and this signal was much stronger and

25. After 5 days of induction with 1mM IPTG at 4 °C, the 424 amino acids recombinant protein accumulated (arrow) in the insoluble fraction (not shown). B. Protein pellet was washed sequentially with TRIS 50mM pH 8, 150mM NaCl and Triton 0.1%, TRIS 50mM pH 8 and 150mM NaCl and finally resuspended in 10mL TRIS 50mM pH 8, NaCl 150mM and N-Lauryl-Sarcosyne (3%) overnight at 4 °C. After dialysis, the total amount of recombinant protein was determined with the Biorad Protein assay kit and concentration was checked on an acrylamide gel by comparison with a known amount of BSA.

26. Five days old seedlings were incubated in pentane for 10' and fixed in formaldhehyde 1.5% - glutaraldhehyde 0.5% in PBST (Triton 0.05% in PBS 1X) overnight at 4 °C followed by 1h under vacuum. After being rinced in PBST (5', 3 times) seedling have been incubated for 1h in pectolyase 0.05% - cellulase 0.05% - 0.4M mannitol in PBS1X at 30 °C, rinced in 0.4M mannitol - PBS1X and then incubated in driselase 2.5% at 37 °C for 45'. Seedlings were then incubated sequentially in PBS1X, methanol (−20 °C, 10'), PBS1X (5', 2 times), glycine 50mM - BSA 1% - PBS1X (30'), glycine 50mM - PBS1X (5'), glycine 50mM - triton 0.5% - PBS1X (2h), glycine 50mM - PBS1X (5'), primary *Totoro* rabbit polyclonal antibody (1/100) - glycine 50mM - PBS1X (4 °C, overnight), glycine 50mM - PBS1X (15', 3 times), goat anti-rabbit AlexaR 488 or 568 antibody (1/200) - glycine 50mM - PBS1X (37 °C, 3h) and PBS1X (15', 3 times).

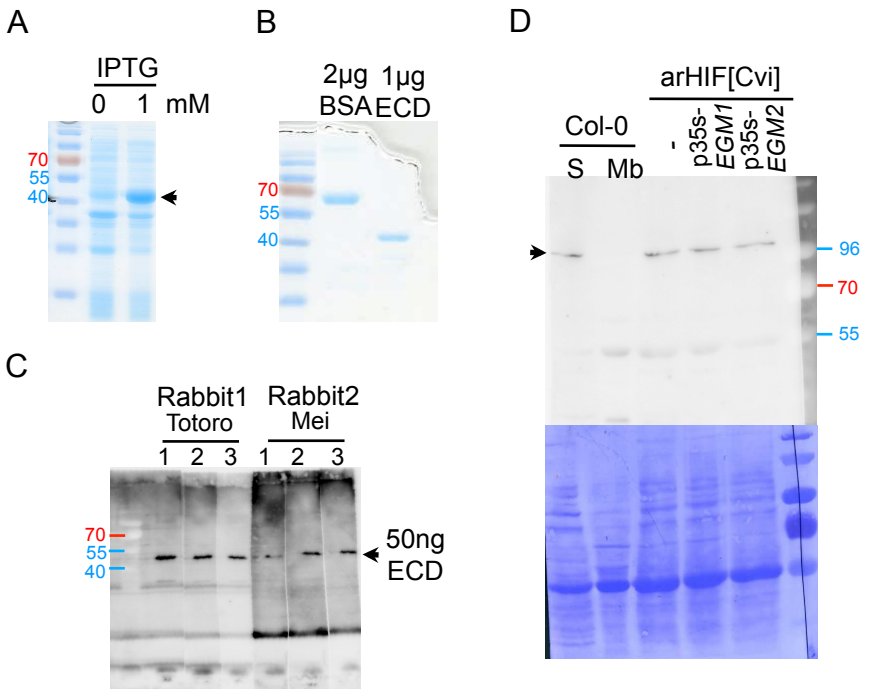


Fig. 36. Production of EGM2 rabbit polyclonal antibody A. The extracellular domain of EGM2 (ECD) (without the signal peptide; 1.206kb) was fused to a 6xHis Tag and produced in *E. coli*. B. The total amount of recombinant protein produced and its relative purity was checked on an acrylamide gel by comparison with a known amount of BSA. C. The two rabbit polyclonal antibodies (*Totoro* and *Mei*) raised against the recombinant protein (Eurogentec) recognised 50 ng of the recombinant EGM2 extracellular domain (arrow) by western blot. Dilution of the serum 1/10 (1), 1/100 (2), 1/1,000 (3), D. A western blot performed on plant extracts using *Totoro* antibody revealed a signal of the expected size (arrow) in all the genotypes tested (Col-0, arHIF[Cvi] and overexpressors– *p35s-EGM1* and *p35s-EGM2*–). However after ultracentrigugation at 100,000g for 30 min, the signal is observed in the soluble fraction (S) and not in the membranous fraction (Mb). Besides, no difference in the intensity of the signal was observed between the arHIF[Cvi] and the overexpressors despite similar protein loading (coomassie). Thus the observed signal is likely to be aspecific.

continuous than the dotted signal observed using preimmune serum (figure 37A-B). Although the signal was likely specifically internalised after mannitol treatment and not sorbitol treatment (figure 37C) and that internalisation was not a general response of membranous proteins under mannitol (figure 37D) we could not confirm the specificity of the antibody in Col-0 plants compared to Col-0 seedlings expressing an amiRNA targeting both *EGM1* and *EGM2* nor in the overexpressors compared to WT arHIF[Cvi] lines (figure 37A-B). Besides, the presence of the signal at the root tip in Col-0 and amiRNA lines was puzzling as all GUS constructs except one *pEGM1:GUS* (figure 37E) show that *EGM1* and *EGM2* are likely expressed in apical meristems and young leaves. Thus, here again the observed signal was likely aspecific. As a conclusion we did not manage to localise EGM1 and EGM2 and so could not test their internalisation under mannitol stress. Anyway, biochemical *in vitro* analyses, such as dot blot, Isothermal Titration Calorimetry [331], Surface Plasmon Reasonance [332], would obviously be necessary to prove the direct link between those RLKs and mannitol, if it exists. Under *in vitro* mannitol treatment, because of unspecific polyol transporter [333, 334], we expect mannitol to enter in the cells and be transported in whole plant organs. As such, because of the relatively restricted expression pattern of *EGM1* and *EGM2* in meristems and young leaves under *in vitro* mannitol stress, it is also possible that these two RLKs do not sense mannitol itself but actually recognize a downstream signal of mannitol stress response or a mannitol-derived compound. Although mannitol is not known to be metabolised in *Arabidopsis thaliana*, such derived compound (1-O-β-D-glucopyranosyl-D-mannitol) has been isolated in *A. thaliana* plants over-expressing the celery *mannose-6-phosphate reductase* gene and accumulating a significant amount of mannitol [297, 334]. Overall, the exact signal perceived by EGM1 and EGM2 still needs to be determined.

Finally, we show that *EGM1* and *EGM2* are two closely related but non-redundant paralogs that likely participate to a common function. They are both induced under mannitol stress in apical meristems and relatively young leaves but *EGM1* also show patches of expression in older leaves (figure 38). It is worth noting that overall, the signal in the *pEGM1:GUS* lines was stronger than the one observed in the *pEGM2:GUS* lines which is consistent with the lower level of expression of *EGM2* compared to *EGM1* observed in the endogenous promoter complementation lines. However during the correction of this manuscript, I realized that a stop codon have been inserted between the ATG of *EGM2* and the GUS reporter in the *pEGM2:GUS* construction, which could strongly result in the differences observed between *EGM1* and *EGM2* expression pattern. In the correct *pEGM1:GUS* lines, the induction of *EGM1* was higher compared to the native copies in the WT line suggesting that a negative regulatory element was probably missing in the 1kb promoters (publication; Figure 3A). Considering the pattern of expression observed with GUS staining, this regulatory element could either restrict further the expression pattern or reduce the cellular expression level. Desipte the error in *pEGM2:GUS*, the overlap between the expression patterns of the two genes suggests that those RLKs could act in

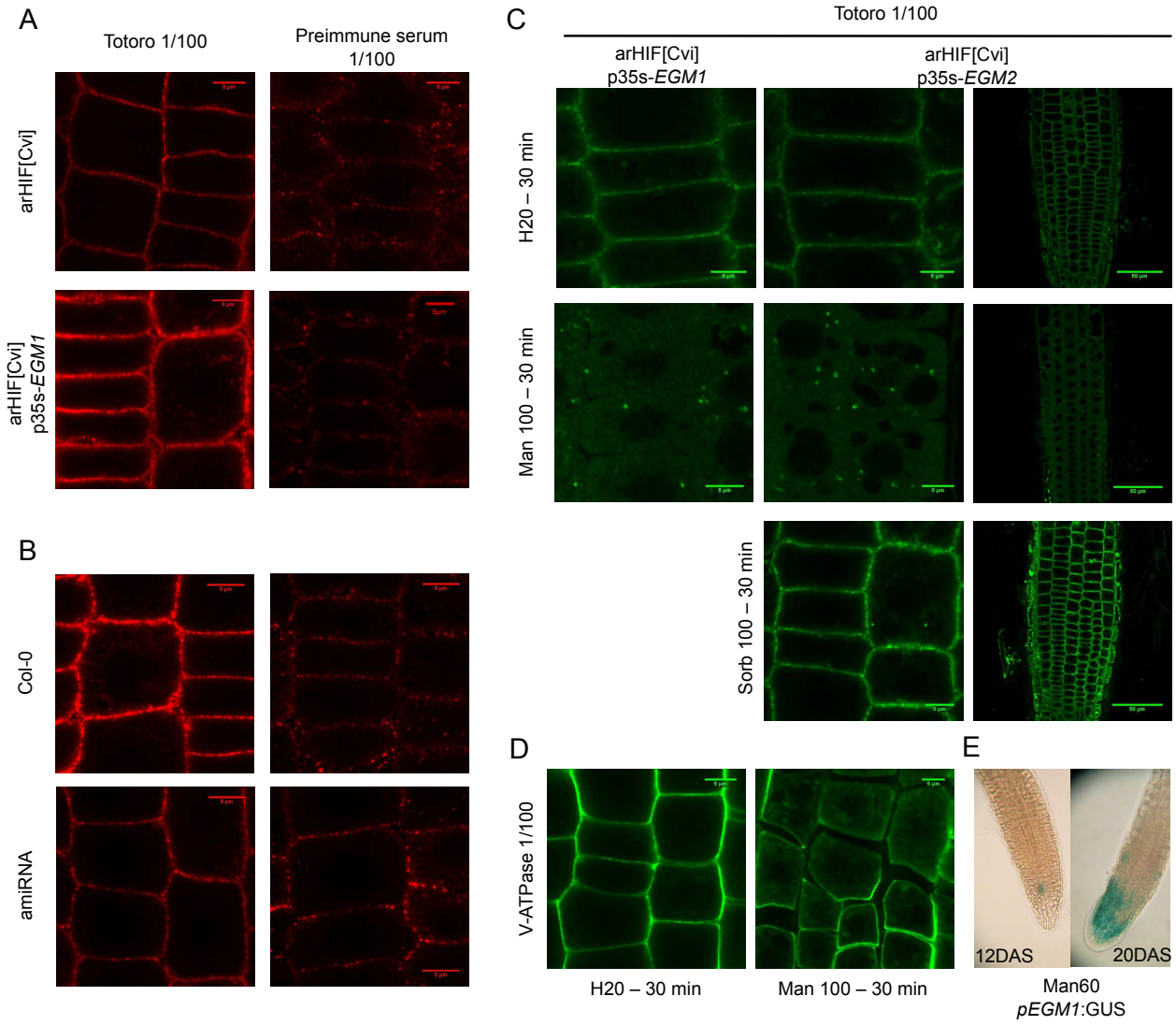


Fig. 37. Immunolocalisation using *Totoro* rabbit polyclonal antibody in apical root cells. The likely membranous signal observed in arHIF[Cvi], overexpressors, Col-0 and amiRNA lines using *Totoro* rabbit polyclonal antibody was much more stronger and continuous than the dotted signal observed using preimmune serum (A-B). Although the signal is likely specifically internalised after mannitol treatment and not sorbitol treatment (C) we could not confirm the specificity of the antibody in mutants nor overexpressors (A-B). Neverthless, the internalisation was specific to *Totoro* antibody as V-ATPase was not internalised after mannitol treatment (D). A, B and C-D correspond to 3 independent experiments. No clear signal was observed in differentiated roots, hypocotyls, leaves, cotyledons nor apical meristems. (E) One *pEGM1:GUS* insertion line out of 9 show coloration at the root tip after mannitol treatment. Signal was observed under ZEISS 710 confocal microscope

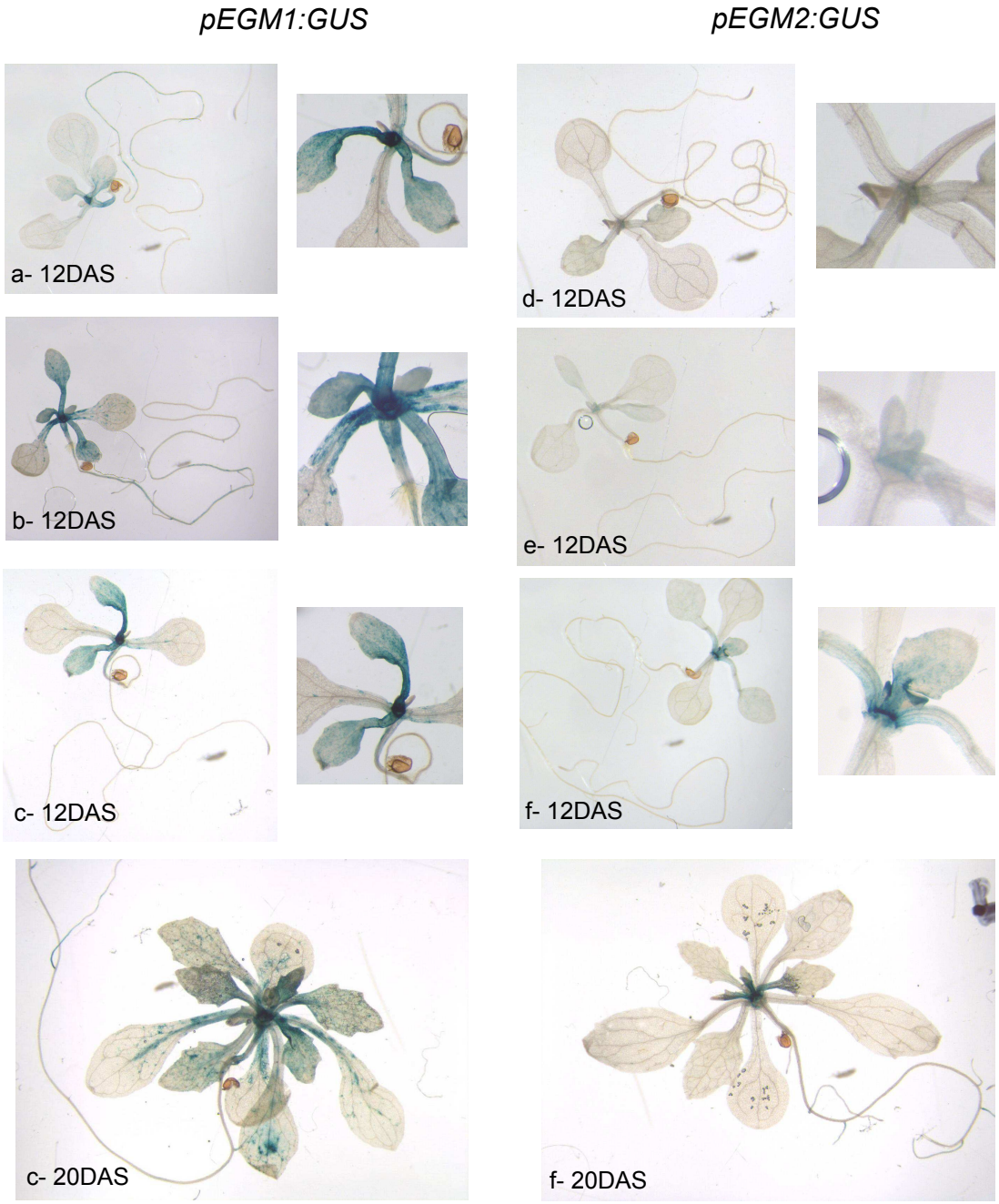


Fig. 38. Pattern of expression of *EGM1* and *EGM2* genes. Histochemical analysis of GUS reporter gene expression driven by *EGM1* or *EGM2* promoter (1kb upstream ATG) in 12DAS or 20DAS whole seedlings grown under Man60 conditions. Three different insertion lines per construct are presented (*pEGM1:GUS* a-c; *pEGM2:GUS* d-f).

the same protein complex. This hypothesis is consistent with the presence of a Pan/apple-like domain involved in protein-protein interaction [316, 335]. Besides, the formation of complexes is a common characteristic of RLK receptors and has been suggested or demonstrated for the most studied RLK such as CLV1/CLV1 [336, 337], FLS2/BAK1 [338, 339], and SRK/SLG (S-RECEPTOR KINASE/S-LOCUS GLYCOPROTEIN) [340]. Conversely, EGM1 and EGM2 could act sequentially in the same pathway. Because the induction of *EGM2* is completely dependent on the functionality of *EGM1* whereas induction of *EGM1* only partially depends on the functionality of *EGM2*, EGM1 would be more likely to act upstream EGM2 under the sequential model. To my mind, because EGM1 could partially substitute EGM2 (publication, Figure2B and 3A) when highly expressed, the model under which EGM1 and EGM2 interact in the same complex is more likely but remain to be proven. It would be interesting to test such interaction using yeast two-hybrid, bi-molecular fluorescence (BIFC) or immuno co-precipitation. Besides, we could have tested whether EGM2 can substitute EGM1 by *egm1* mutants transformed with *pEGM2:EGM2* construct.

10.3 EGM1 and EGM2 natural variation and evolution.

Beyond our hypothesis that EGM receptors are associated with pathogen defence, an interesting question might be the evolution and adaptive potential of those SD1-RLKs. First, the mutations we identified in *EGM2[Cvi]* as responsible for enhanced shoot growth under mannitol stress are found in 13 other accessions (over around 500) that are relatively widespread in Eurasia (figure 39A). Considering the relatively low frequency of these mutations and their likely deleterious effect associated with increased pathogen sensitivity, we could hypothesize that they appear under simple genetic drift or slightly counter-selected but maintained due to demographic constraints (selfing and/or low population size). Another possibility could be that these two RLKs are under balancing selection so that part of the alleles are functional and others not. This type of selection is relatively frequent for defence-related genes because of fitness cost in the absence of pathogens [264, 265, 261, 262, 263]. In our case, plants with functional *EGM1* and *EGM2* Col-0 alleles did not show biomass reduction neither *in vitro* nor *in vivo* compared to plants with Cvi-0 hypofunctional allele. Besides, at the worldwide scale in 408 accessions, neither *EGM1* nor *EGM2* Tajima's D values and Fay and Wu's H values are consistent with balancing selection (table 6). Nevertheless, the haplotype network obtained from the alignment of the region covering *EGM1* and *EGM2* in 408 accessions shows that some haplotypes accumulated many mutations in the extracellular domain of *EGM1* (haplotypic cluster a), in the kinase domain of *EGM1* and promoter of *EGM2* (haplotypic cluster b) compared to other haplotypes (cluster c) (figure 39B-C). In our paper, we show that Sha EGM allele, that belongs to the (b) haplotypic cluster, is likely functional compared to Cvi

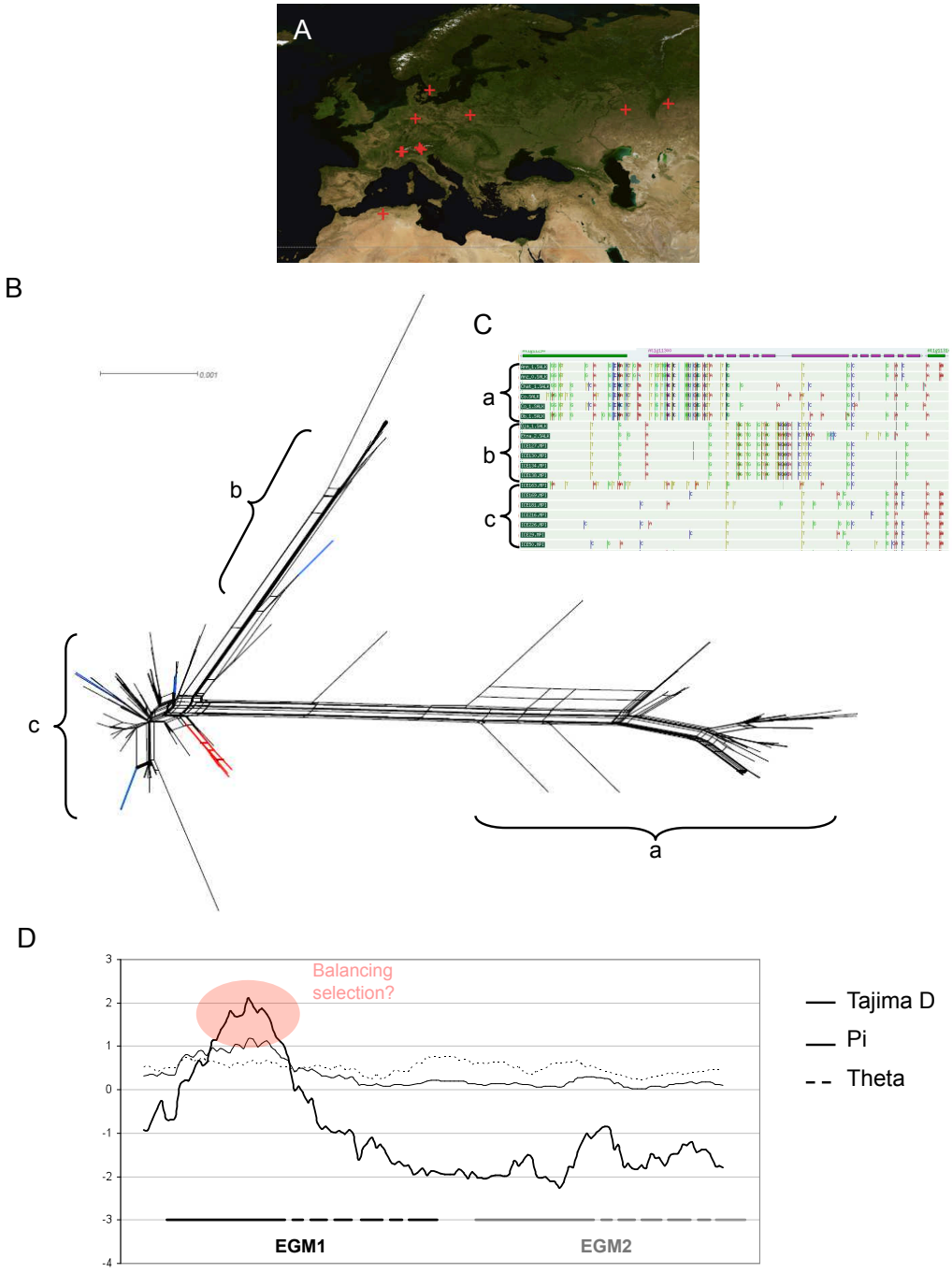


Fig. 39. Genetic diversity of *EGM1* and *EGM2* RLKs A. Distribution of the 13 accessions with an *EGM2[Cvi]*-like allele. Cvi-0 is not represented on the map. B. Neighbornet haplotype network obtained using an alignment of 408 alleles of the whole *EGM* region (including *EGM1* and *EGM2*). Red edges correspond to the *EGM2[Cvi]*-like alleles. Blue edges correspond to 4 of the Col-like accessions used in specific association genetics approach. C. Screen shots of Salk G-browse representing the main alleles observed in *EGM1* and *EGM2* region. D. Sliding-window analysis of nucleotide diversity (π and θ x100 for scale issues) and Tajima's D in 408 accessions at *EGM* locus. Window size is 500 bp, and step size is 50bp. The position of *EGM1* (black) and *EGM2* (grey) exons is indicated.

Tab. 6. Genetic diversity at *EGM*1 and *EGM2*.

		N (a)	Length	S (b)	π (c)	θ_w (d)	Divergence (e)	Tajima's D (f)	Fay & Wu (g)	Divergence (h)
EGM1	Worldwide	408	3365	113	0.0049	0.00514	0.0498	-0.14 (0.53)	-15.40 (0.055)	0.125
EGM2	Worldwide	408	3355	103	0.0013	0.00473	0.119	-2.13 (0.001)	-24.40 (0.016)	-

(a) Number of sequences – (b) Number of segregating sites in the region – (c) Number of nucleotide differences per site between two sequences – (d) Theta per site from S – (e) Average number of nucleotide substitution per site between *A. thaliana* and A. lyrata – (f) Tajima's D statistics value (p-value simulated from 1000 replicates using coalescent theory) – (f) Fay and Wu statistics value (p-value simulated from 1000 replicates using coalescent theory) – (h) Average number of nucleotide substitution per site between EGM1 and EGM2

EGM allele, which suggests that members of cluster (b) might be functional alleles. However, we did not test accessions belonging to cluster (a). This cluster is particularly interesting as it seems responsible for an increase in Tajima's D value within the extracellular domain of *EGM1* which could be interpreted as traces of balancing selection (figure 39D). Further functional and population genetics analyses are required to properly conclude on the evolutionary potential of these two RLKs. Besides, it could be interesting to analyse in more details the diversity and trace of selections in the whole SD1 family that is likely enriched in genes responding to biotic stress [319].

Part IV

NATURAL EPIGENETIC VARIATION AT *QQS LOCI* IN *A. THALIANA*

11. PROJECT BACKGROUND AND PERSONAL CONTRIBUTION

Michel Vincentz and his PhD student, Amanda Silveira, are the main investigators for this project. While they were studying a transcription factor possibly involved in energy homeostasis, they observed that *QQS* (for qua-quine starch), a gene known to be involved in starch metabolism [341], was down regulated in a mutant of this transcription factor compared to WT plants. Further analyses revealed that *QQS* was actually upregulated in the WT line they were using (a background descending from the Col-0 accession) and that this had no relation with any T-DNA insertion. They called this WT line 'Col-0*'. Using a combination of genetic and epigenetic approaches they had showed that different stable epialleles of the *QQS* gene were segregating in Col-0* seed stock as well as in some *A. thaliana* accessions. Those epialleles were characterized by a strong negative correlation of the methylation level of *QQS* promoter and 5' UTR with *QQS* transcript accumulation. From those results they asked us to participate in different aspects of the project.

First, in the lab, several people are working on identifying and confirming expression quantitative trait loci (eQTL) [27] in Cvi-0 x Col-0 and Bur-0 x Col-0 RILs populations. Among all the local-eQTL identified, *QQS* was one of the most significant segregating in the Cvi-0 x Col-0 RIL set [181]. Using the expertise of the lab, and with the precious help of Matthieu Canut, we confirmed that *QQS* was indeed controlled in cis by this eQTL, by performing allele-specific expression tests [28] in Jea x Col-0 and Kond x Col-0 hybrids (Jea and Kond being both demethylated at *QQS*, like Cvi) (Publication – Fig. S4F). Besides, we show that we can restore *QQS* expression in methylated accessions (Col-0 and Shahdara) by treating plants with a methylation inhibitor (Publication – Fig. S4D, Fig. S4F). Then, because Olivier collected several accessions from Central Asia during his career and had some remaining seeds collected directly in the field, we were able to test whether *QQS* epialleles could really segregate in natural en-

27. eQTL are loci that differently regulate the level of expression of one or several genes in different *A. thaliana* accessions. Those loci can be either distant (distant eQTL) or close (local eQTL) to the gene they are regulated. An eQTL located in the regulatory elements of the target gene and that affects specifically the transcript abundance of the linked allele is named cis eQTL whereas eQTLs that are located elsewhere and affect the transcript abundance of both alleles of the target genes are named trans eQTLs.

28. Allele-specific expression (ASE) assays consist in quantifying the level of each allele in F1 (hybrid) mRNA in comparison to 1:1 parental mRNA pools. If the ratio of each allele's expression is different from 1 in the hybrid it means that this differential regulation is acting in cis. Moreover, if this ratio is the same in the hybrid and in the parental pool, then it means that there is no additional trans contribution to the difference in allelic expression.

vironments and were not the result of multiple selfing generations in experimental greenhouse conditions (Publication – Fig. 4B, Fig. 4C). Finally, to analyse the adaptive potential of *QQS* epialleles and because the lab is used to perform high-throughput phenotyping, we tried to find a phenotype that differentiate the different epialleles (figure 40, figure 41, figure 42).

12. PUBLICATION IN PLOS GENETICS

EXTENSIVE NATURAL EPIGENETIC VARIATION AT A DE NOVO ORIGINATED GENE.

The *QQS* story has been published in PLoS Genetics in April 2013. My contributions to this paper are highlighted in figures 4B, 4C, S4D and S4F.

OPEN ACCESS Freely available online

PLOS | GENETICS

Extensive Natural Epigenetic Variation at a *De Novo* Originated Gene

Amanda Bortolini Silveira[1], Charlotte Trontin[2], Sandra Cortijo[3], Joan Barau[1], Luiz Eduardo Vieira Del Bem[1], Olivier Loudet[2], Vincent Colot[3]*, Michel Vincentz[1,4]*

1 Centro de Biologia Molecular e Engenharia Genética, Universidade Estadual de Campinas, Cidade Universitária "Zeferino Vaz," Campinas, São Paulo, Brazil, **2** INRA, UMR 1318, Institut Jean-Pierre Bourgin, Versailles, France, **3** Ecole Normale Supérieure, Institut de Biologie de l'ENS (IBENS), Centre National de la Recherche Scientifique (CNRS) UMR 8197, Institut National de la Santé et de la Recherche Médicale (Inserm) U1024, Paris, France, **4** Departamento de Biologia Vegetal, Universidade Estadual de Campinas, Cidade Universitária "Zeferino Vaz," Campinas, São Paulo, Brazil

Abstract

Epigenetic variation, such as heritable changes of DNA methylation, can affect gene expression and thus phenotypes, but examples of natural epimutations are few and little is known about their stability and frequency in nature. Here, we report that the gene *Qua-Quine Starch* (*QQS*) of *Arabidopsis thaliana*, which is involved in starch metabolism and that originated *de novo* recently, is subject to frequent epigenetic variation in nature. Specifically, we show that expression of this gene varies considerably among natural accessions as well as within populations directly sampled from the wild, and we demonstrate that this variation correlates negatively with the DNA methylation level of repeated sequences located within the 5′end of the gene. Furthermore, we provide extensive evidence that DNA methylation and expression variants can be inherited for several generations and are not linked to DNA sequence changes. Taken together, these observations provide a first indication that *de novo* originated genes might be particularly prone to epigenetic variation in their initial stages of formation.

Citation: Silveira AB, Trontin C, Cortijo S, Barau J, Del Bem LEV, et al. (2013) Extensive Natural Epigenetic Variation at a *De Novo* Originated Gene. PLoS Genet 9(4): e1003437. doi:10.1371/journal.pgen.1003437

Editor: Michael D. Purugganan, New York University, United States of America

Received December 14, 2012; **Accepted** February 21, 2013; **Published** April 11, 2013

Funding: This work was supported by grants from Fundação de Amparo a Pesquisa do Estado de São Paulo (FAPESP - BIOEN Program) to MV, CNRS/FAPESP to VC and MV, and the European Union Network of Excellence Epigenesys to VC. ABS was supported by a PhD studentship from FAPESP, and CT and SC by PhD studentships from the French Ministry of Research. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: colot@biologie.ens.fr (VC); mgavince@unicamp.br (MV)

Introduction

DNA mutations are the main known source of heritable phenotypic variation, but epimutations, such as heritable changes of gene expression associated with gain or loss of DNA methylation, are also a source of phenotypic variability. Indeed, several stable DNA methylation variants affecting a wide range of characters have been described, mainly in plants [1–3]. In most instances, epimutations are linked to the presence of structural features near or within genes, such as direct [4–6] or inverted repeats [7,8] or transposable element (TE) insertions [9], which act as units of DNA methylation through the production of small interfering RNAs (siRNAs) [3,10]. Examples of epimutable loci in *Arabidopsis thaliana* (*A. thaliana*) include the *PAI* [7] and *ATFOLT1* genes [8], which have suffered siRNA-producing duplication events in some accessions and also the well characterized *FWA* locus, which contains a set of SINE-derived siRNA-producing tandem repeats at its 5′end [4,5]. Repeat-associated epimutable loci are almost invariably found in the methylated form [5–9] in nature, which reflects, at least in part, that DNA methylation is particularly well-maintained over repeats [11,12]. Indeed, epigenetic variation at *PAI*, *ATFOLT1* and *FWA* has only been observed in experimental settings. Similarly, sporadic gain or loss of DNA methylation associated with changes in gene expression has only been documented in *A. thaliana* mutation accumulation lines [13,14] and examples of natural epigenetic variation in other plant species are few [15–17].

Here we report a case of prevalent natural epigenetic variation in *A. thaliana*, which concerns a *de novo* originated gene [18]. We show that expression of this gene, named *Qua-Quine Starch* (*QQS*), is inversely correlated with the DNA methylation level of its promoter and that these variations are stably inherited for several generations, independently of DNA sequence changes. Based on these findings, we speculate that epigenetic variation could be particularly beneficial for newly formed genes, as it would enable them to explore more effectively the expression landscape than through rare DNA sequence changes.

Results

QQS Is a Novel Gene Embedded within a TE-Rich Region of the *A. thaliana* Genome and Is Negatively Regulated by DNA Methylation

The *A. thaliana Qua-Quine Starch* (*QQS*, *At3g30720*) was first described as a gene involved in starch metabolism in leaves [19,20]. Despite being functional and presumably already under purifying selection ($dN/dS = 0.5868$; p-value<0.045), *QQS* is likely a recent gene that emerged *de novo*. Indeed, *QQS* has no significant similarity to any other sequence present in GenBank [18,19],

Author Summary

Epigenetics is defined as the study of heritable changes in gene expression that are not linked to changes in the DNA sequence. In plants, these heritable variations are often associated with differences in DNA methylation. So far, very little is known about the extent and stability of epigenetic variation in nature. In this study, we report a case of extensive epigenetic variation in natural populations of the flowering plant *Arabidopsis thaliana*, which concerns a gene involved in starch metabolism, named *Qua-Quine Starch* (*QQS*). We show that in the wild *QQS* expression varies extensively and concomitantly with DNA methylation of the gene promoter. We also demonstrate that these variations are independent of DNA sequence changes and are stably inherited for several generations. In view of the recent evolutionary origin of *QQS*, we speculate that genes that emerge from scratch could be particularly prone to epigenetic variation. This would in turn endow epigenetic variation with a unique adaptive role in enabling *de novo* originated genes to adjust their expression pattern.

suggesting that it originated from scratch since *A.thaliana* diverged from its closest sequenced relative *A. lyrata* around 5–10 million years ago. Furthermore, *QQS* encodes a short protein (59 amino acids) and it is differentially expressed under various abiotic stresses [18], which are also hallmarks of *de novo* originated genes [21–23].

As shown in Figure 1, *QQS* is surrounded by multiple transposable element sequences (Figure 1A) and contains several tandem repeats in its promoter region and 5′UTR (Figure 1B). In the Columbia (Col-0) accession, these tandem repeats are densely methylated and produce predominantly 24 nt-long siRNAs (Figure 1B, Figure S1A and S1B). Publically available transcriptome data [24,25] and results of RT-qPCR analyses (Figure S1C) show that steady state levels of *QQS* mRNAs are higher in several mutants affected in the DNA methylation of repeat sequences, including *met1* (*DNA METHYLTRANSFERASE 1*), *ddc* (*DOMAINS REARRANGED METHYLTRANSFERASE 1 and 2 and CHROMOMETHYLASE 3*), *ddm1* (*DECREASE IN DNA METHYLATION 1*) and *rdr2* (*RNA-DEPENDENT RNA POLYMERASE 2*), which abolishes the production of 24 nt-long siRNAs as well as most CHH methylation. These findings indicate that *QQS* expression is negatively controlled by DNA methylation and point to the siRNA-producing tandem repeats as being potentially involved in this repression.

Stably Inherited Spontaneous and Induced Epigenetic Variation at *QQS*

We first observed epiallelic variation at *QQS* unexpectedly, in a Col-0 laboratory stock (hereafter referred to as Col-0*) with increased expression of the gene and decreased DNA methylation of its promoter and 5′UTR repeat elements (Figure 2A). No sequence change could be detected in the Col-0* stock within a 1.2 kb region covering the *QQS* gene (Figure 1B), which excluded local *cis*-regulatory DNA mutations at the *QQS* locus as being responsible for DNA methylation loss in Col-0*. Additionally, comparative genomic hybridization analysis as well as genome-wide DNA methylation profiling using methylated DNA imunoprecipitation assays revealed no major differences between Col-0 and Col-0* (Figure S2).

We next investigated the *QQS* epigenetic status in pooled seedlings (S1) derived from the selfing of 12 individual Col-0* plants (Figure S3). Results revealed a range of *QQS* epialleles and a strong negative correlation between DNA methylation and expression of the gene (Figure 2B and 2C). To explore further this variation, a single S1 individual was then selfed for each of the 12 lines and seedlings (S2) were analyzed in pool for each line, as above (Figure S3). Remarkably, the differences in *QQS* expression and DNA methylation observed at the S1 generation were also observed at the S2 generation (Figure 2B and 2C). Taken together, these results suggest therefore the existence of a range of epiallelic variants at *QQS*, which are stably inherited for at least one generation.

The inheritance of *QQS* hypomethylated epialleles was also examined in a random sample of 19 *ddm1*-derived epigenetic Recombinant Inbred Lines (epiRILs) obtained by crossing a Col-0 wild-type (wt) line with an hypomethylated Col-0 *ddm1* line [26]. High DNA methylation/low expression and low DNA methylation/high expression of *QQS* were observed in 14 and 5 epiRILs, respectively (Figure 2D). This is consistent with Mendelian segregation of the highly methylated/lowly expressed Col-0 wt and lowly methylated/highly expressed Col-0 *ddm1* parental *QQS* epialleles (75%/25% expected because of backcrossing rather than selfing of the F1; $Chi^2 = 0{,}017$, p-value$>$0.05). Indeed, examination of the epi-haplotype obtained for 17 of these epiRILs [27] confirmed the wt or *ddm*-origin of the *QQS* locus in each case (data not shown). These results demonstrate therefore that, like many other *ddm1*-induced epialleles [28,29], *QQS* hypomethylated epialleles can be stably inherited for at least eight generations and are not targets of paramutation.

QQS Is under Autonomous Epigenetic Control

We next investigated the degree to which DNA methylation of *QQS* and of flanking TEs are independent from each other. To this end, we first analyzed DNA methylation patterns of TE sequences flanking *QQS* in a series of epiRIL with contrasted *QQS* epialleles. Unlike for *ddm1*-derived *QQS*, hypomethylation was not inherited for the three TEs located immediately upstream of the gene, as they did systematically regain wt DNA methylation levels (Figure 3A and 3B), presumably because of their efficient targeting by RNA-directed DNA methylation (RdDM) [28]. In addition, although the TE just downstream of *QQS* was always hypomethylated when inherited from *ddm1*, hypomethylation was also observed in one epiRIL that inherited the *QQS* region from the wt parent. Thus, there is no strict correlation between DNA methylation at *QQS* and this downstream TE. We next examined the effect of several T-DNA and transposon insertions located ~3.1 kb or 153 bp upstream of the transcription start site (TSS), 653 bp downstream of the 3′UTR and within the second coding exon of *QQS*. Whereas three of these insertions had no effect on DNA methylation and expression levels of *QQS*, the T-DNA insertion located closest to the TSS was associated with a drastic reduction of DNA methylation of both the promoter and 5′UTR of the gene, as well as with an increase in *QQS* expression (Figure 3A and 3C). However, this insertion had no impact on DNA methylation of upstream and downstream TEs (Figure 3A and 3D). Taken together, these results suggest that epigenetic variation at *QQS* is most likely determined by sequences within the promoter and 5′UTR of the gene, not by the TEs that are located immediately upstream or downstream.

QQS Exhibits Epigenetic Variation among Natural Accessions

We next investigated the possibility that *QQS* is subject to epigenetic variation in natural populations. To this end, we first analyzed *QQS* expression and DNA methylation in 36 accessions

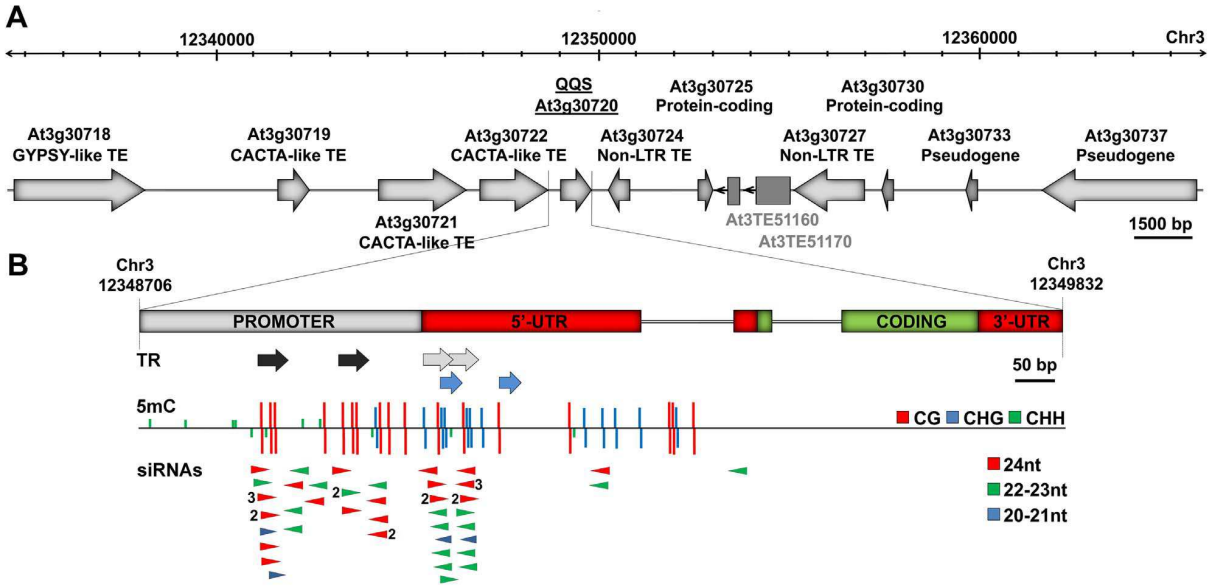


Figure 1. *QQS* is embedded in a repeat-rich region. (A) Genomic structure of the *QQS* locus (30 kb window) in the Col-0 accession. Dark grey boxes represent two additional TE sequences predicted by [51,52]. (B) Magnified view of the *QQS* gene and upstream sequences, showing tandem repeats (TR), methylation of cytosine residues (5 mC) at the three types of sites (CG, CHG and CHH, H = A, T or C) and locus-specific sense and antisense siRNAs (numbers referring to copy number). DNA methylation and siRNA data are from [25].
doi:10.1371/journal.pgen.1003437.g001

representing the worldwide diversity [30]. *QQS* was methylated and lowly expressed in 29 accessions, but unmethylated and highly expressed in seven accessions distributed over the entire geographic range (Figure 4A). This indicates that epigenetic variation at *QQS* is widespread in nature. In contrast, upstream and downstream TEs were consistently methylated in all accessions (Figure S4A and S4B), thus confirming that the epigenetic state at *QQS* is not determined by that of flanking TEs. We then sequenced a 2.8 kb interval encompassing the *QQS* gene and its flanking regions from the seven accessions carrying the hypomethylated/highly expressed epiallele as well as from three accessions carrying a methylated/lowly expressed epiallele. Although several SNPs and indels were identified (Figure S4C), no correlation between any specific sequence alterations and *QQS* DNA methylation or expression states could be established (Figure 4A). In addition, while Kondara and Shahdara have identical *QQS* sequences, they have contrasted DNA methylation/expression patterns at the locus (Figure 4A and Figure S4C), which provides further evidence that natural epiallelic variation at *QQS* is independent of local *cis*-DNA sequence polymorphisms and is thus most likely truly epigenetic. Analysis of a Cvi-0 vs. Col-0 Recombinant Inbred Line (RIL) population revealed in addition that *QQS* expression is controlled by a large-effect local-expression quantitative trait locus (local-eQTL; http://qtlstore.versailles.inra.fr/) [31]. This suggests that like the Col-0 wt and Col-0 *ddm1 QQS* epialleles (Figure 2D), the Cvi-0 hypomethylated *QQS* epiallele is stably inherited across multiple generations. This further demonstrates that epigenetic variation at *QQS* is not appreciably affected by sequence or DNA methylation polymorphisms located elsewhere in the genome and indicates also that *QQS* is not subjected to paramutation [29].

To validate experimentally the causal relationship between DNA methylation and repression at *QQS*, seedlings of several accessions were grown in the presence of the DNA methylation inhibitor 5-aza-2′-deoxycytidine (5-aza-dC). In the two accessions Col-0 and Shahdara, which harbor distinct methylated and lowly expressed *QQS* alleles, treatment resulted in reduced DNA methylation and increased expression of *QQS* (Figure S4D). In contrast, seedlings of Jea, Kondara and Cvi-0 accessions, all of which harbor a demethylated/highly expressed *QQS* allele, did not show further reduction of DNA methylation or increased expression when grown in the presence of the demethylating agent (Figure S4D). Moreover, whereas expression of *QQS* in F1 hybrids derived from crosses between Col-0 (methylated *QQS*) and Kondara (hypomethylated *QQS*), was always higher for the epiallele inherited from the hypomethylated parent, further confirming that *QQS* is not subjected to paramutation [29], treatment with 5-aza-dC reduced dramatically this expression imbalance, most likely as a consequence of demethylation of the Col-0-derived *QQS* allele (Figure S4E). Taken together, these results clearly demonstrate that DNA methylation at *QQS* is causal in repressing expression of the gene.

Wild Populations from Central Asia Exhibit Epigenetic Variation at *QQS*

Finally, we asked whether epigenetic variation at *QQS* could be observed in natural settings or if such variation only emerged in the laboratory, where accessions are grown under controlled growth conditions. To this end, we analyzed *QQS* expression and DNA methylation in plants grown from seeds directly collected from wild populations in Tajikistan, Kyrgyzstan and Iran (NeoShahdara, Zalisky and Anzali populations, respectively). Widespread *QQS* epiallelic variation was observed, both between and within these diverse wild populations (Figure 4B). In addition, *QQS* epigenetic variation was examined in the offspring (after two single seed descent generations) of 25 NeoShahdara individuals. These individuals were randomly sampled among a single patch of several thousands of plants that presumably represent the direct descendants of the Shahdara accession. Based on 10 microsatellite markers and one InDel marker, two genetically distinct

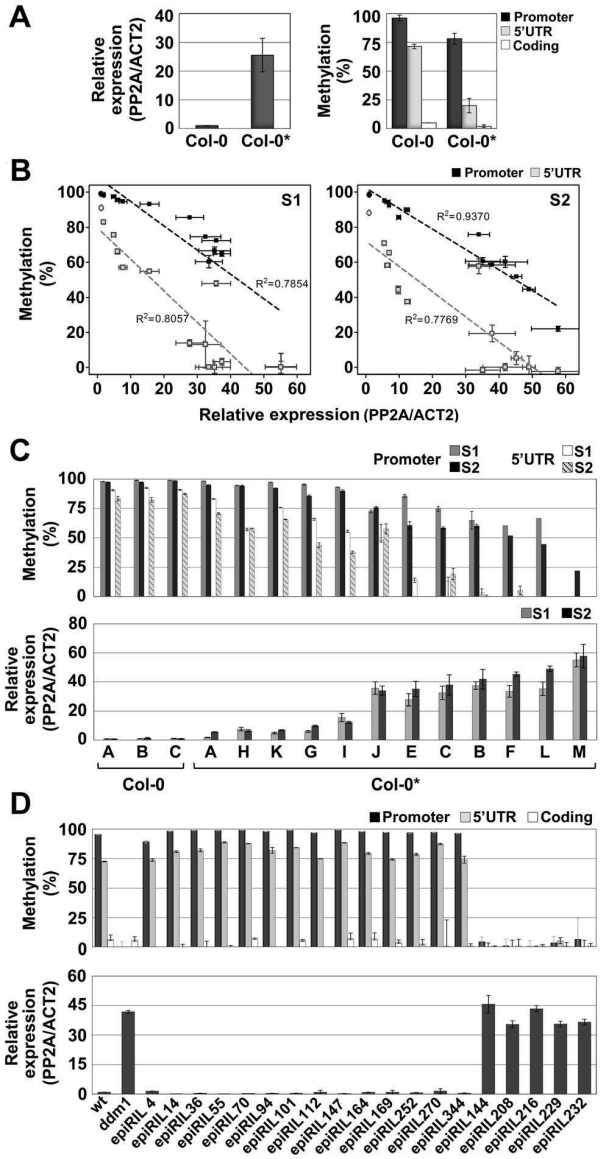


Figure 2. Spontaneous and induced epigenetic variation at *QQS*. (A) DNA methylation and expression profiles of *QQS* in seedlings of the Col-0 and Col-0* stocks. (B) and (C) Negative correlation between *QQS* DNA methylation and expression levels in pooled seedlings of Col-0 and Col-0* (represented by circles and squares in B, respectively) S1 and S2 generation single seed descent lines. (D) DNA methylation and expression levels of *QQS* in seedlings of *ddm1*-derived epiRILs. Error bars represent standard deviation observed in three biological replicates (A–D – expression; A – DNA methylation) or two technical replicates (B–D – DNA methylation).
doi:10.1371/journal.pgen.1003437.g002

subpopulations could be identified. While *QQS* was highly methylated/lowly expressed in all 16 individuals of subpopulation #1, clear differences in DNA methylation and expression were detected among the 9 individuals of subpopulation #2 (Figure 4C). Whether epiallelic variation at *QQS* in subpopulation #2 reflects inherent fluctuations or an intermediary stage in the route to fixation of one of the two epiallelic forms remains to be determined.

Discussion

QQS is a protein-coding gene that likely originated *de novo* in *A. thaliana* within a TE-rich region (Figure 1A). We have shown that this gene, which contains short repeat elements matching siRNAs (Figure 1B, Figure S1A and S1B), varies considerably in its DNA methylation and expression in the wild (Figure 4). We also show that these variations are heritable and independent of the DNA methylation status of neighboring TEs or of DNA sequence variation, either in *cis* or *trans* (Figure 2 and Figure 3, Figures S2 and S4). Thus, we can conclude that *QQS* is a hotspot of epigenetic variation in nature. Consistent with this, *QQS* is among the few genes for which spontaneous DNA methylation variation was observed in Col-0 mutation accumulation lines [13].

Cytosine methylation at *QQS* concerns CG, CHG and CHH sites, which is the pattern expected for sequences with matching siRNAs (Figure 1B, Figure S1B). All three types of methylation sites likely contribute to silencing of *QQS*, as judged by the reactivation of *QQS* in the *met1*, *ddm1*, *ddc* and *rdr2* mutant backgrounds (Figure S1C; [24,25]). Yet, among the different DNA methyltransferases targeting DNA methylation at *QQS*, *MET1* may play a more prominent role, given that DNA methylation at this locus is only fully erased in *met1* mutant plants [25]. *QQS* demethylated epiallelic variants may thus preferentially arise through spontaneous [13] or stress-induced [10] defects in DNA methylation maintenance and be stably inherited for multiple generations as a result of the concomitant loss of matching siRNAs, which would prevent efficient remethylation and silencing of the gene [28,29]. Indeed, although we could not detect *QQS* siRNAs by Northern blot analysis, presumably because of their low abundance, deep sequencing data indicate that they accumulate less in *met1* mutant plants than in wild type Col-0 [25].

Few genes have been shown so far to be subject to heritable epigenetic variation in *A. thaliana* [5–8,13,14,32] and *QQS* is unique among these in exhibiting this type of variation in nature (Figure 4). This therefore raises the question as to what distinguishes *QQS* from other genes, such as *FWA*, for which epigenetic variation can be readily induced in the laboratory in advanced generations of *ddm1* and *met1* mutant plants [5,33], but for which this type of variation is not observed among accessions [11,34]. One possibility is that unlike *QQS* epivariants, *fwa*-hypomethylated epialleles are strongly counter-selected because of their potentially maladapted phenotype, namely late flowering [5]. Consistent with this explanation, epiallelic variation with no phenotypic consequences has been documented at *FWA* in other Arabidopsis species. In these cases, however, inheritance across multiple generations has not been rigorously tested [35]. Another possibility is that *de novo* originated genes, such as *QQS*, are particularly prone to heritable epigenetic variation. This is a reasonable assumption considering that these genes tend to lack proper regulatory sequences initially, unlike new gene duplicates, which by definition come fully equipped [21]. In turn, given that epigenetic variation enables genes to adjust their expression in a heritable manner much more rapidly than through mutation while preserving the possibility for rapid reversion, it could prove particularly beneficial in the case of genes that are created from scratch. Once the most adaptive expression state is reached, it could then become irreversibly stabilized (i.e. genetically assimilated) through DNA sequence changes. Although speculative, this proposed scenario could be highly significant given the recent discovery that *de novo* gene birth may be more prevalent than gene duplication [23].

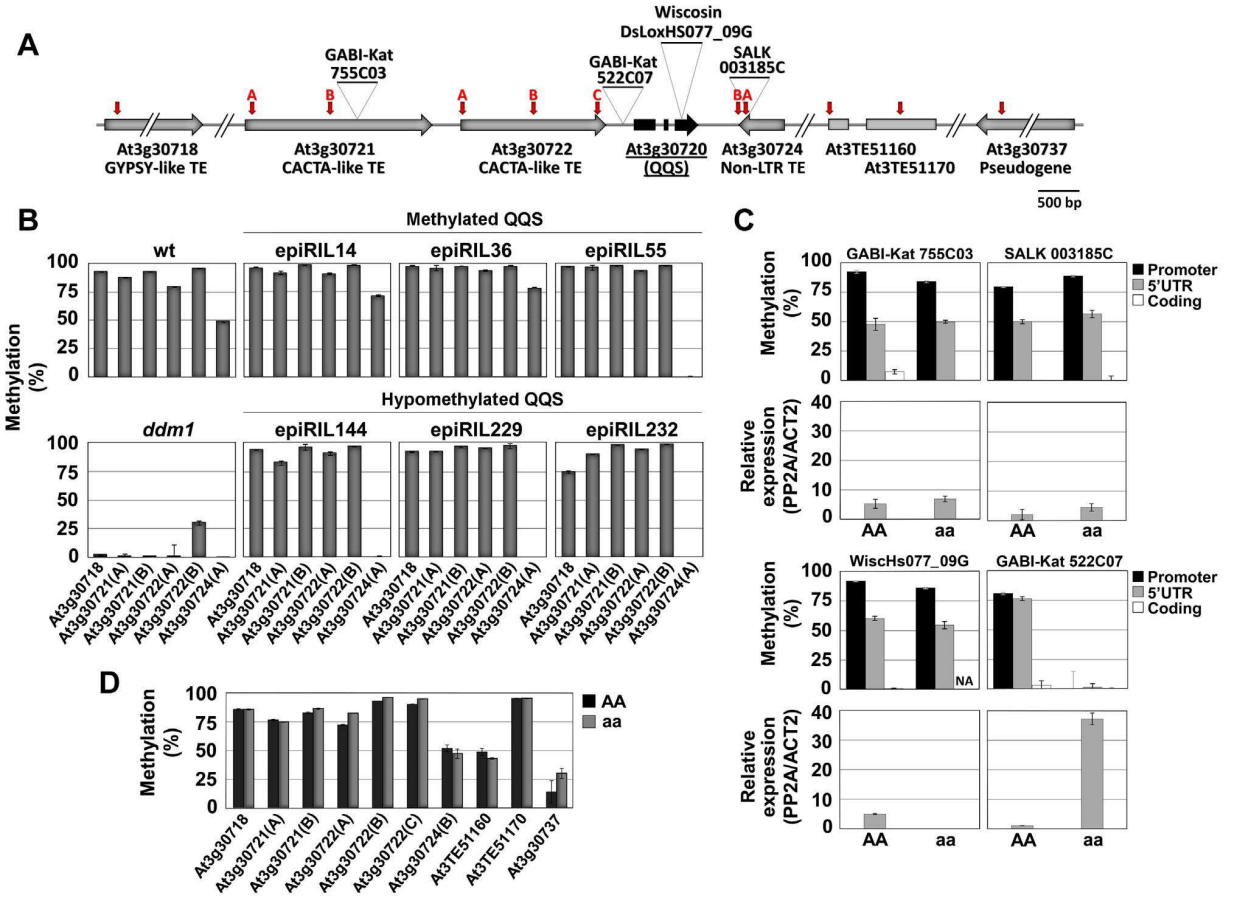


Figure 3. Epigenetic variation at *QQS* is determined by proximal sequences. (A) Schematic representation of the T-DNA/Transposon insertion sites (triangles; GABI-Kat 755C03, GABI-Kat 522C07, WiscDsLoxHs077_09G (WiscHs077_09G) and SALK 003185C) and McrBC-qPCR primer pairs used (vertical arrows; A, B and C represent different primer pairs designed for the same element). (B) DNA methylation levels of TEs flanking *QQS* in epiRILs that had inherited a wt or a *ddm1*-derived *QQS* epiallele. (C) DNA methylation and expression levels of *QQS* in lines carrying the T-DNA/transposon insertions represented in (A). (D) DNA methylation levels of TEs flanking *QQS* in the GABI-Kat 522C07 T-DNA insertion line. AA and aa represent wt and T-DNA homozygous individuals, respectively, coming from the selfing of one hemizygous (Aa) plant. NA, not analyzed. Error bars represent standard deviation observed in two technical replicates (B and D) or three biological replicates (C).
doi:10.1371/journal.pgen.1003437.g003

Materials and Methods

Plant material and growth conditions

A. thaliana accessions were obtained from the INRA Versailles collection (dbsgap.versailles.inra.fr/vnat/, www.inra.fr/vast/collections.htm) [30,36,37]. Insertion lines were obtained from the GABI-Kat at University of Bielefeld, Germany (GABI-Kat 755C03 and 522C07) [38], the ABRC at Ohio State University (SALK 003195C) and University of Wisconsin, Madison, US (WiscDsLoxHs077_09) [39]. Seeds of *ddm1-2* [40], *rdr2-1* [41] and *ddm1*-derived epiRIL lines [26] were provided by V.Colot. NeoShahdara individuals were genotyped with 10 microsatellite markers (NGA8, MSAT2.26, MSAT2.4, NGA172, MSAT3.19, ICE3, MSAT3.1, MSAT3.21, MSAT4.18, ICE5; http://www.inra.fr/vast/msat.php) and one InDel marker in MUM2 gene (MUM2_Del-LP TGGTCGTTATTGGGTCTCGT, MUM2 Del-RP TTAAGAACGCCCGAGGAATA). For expression and DNA methylation assays, seedlings were grown in vitro (MS/2 media supplemented with 0,7% sucrose) for eight days in a culture room (22°C, 16 hours light/8 hours dark cycle, 150 µmol s−1 m−2). Treatment with 5-aza-2′-deoxycytidine was performed as described in [8].

RT–qPCR analysis of *QQS* expression

Total RNA was isolated as described in [42] and cDNA was synthetized using oligo(dT) primers and IMPROM II reverse transcriptase (Promega). Real time PCR reactions were run on an Applied Biosystems 7500 Real-Time PCR System using Platinum SYBR green (Invitrogen). *QQS* expression levels relative to *Actin2/PP2A* or *PP2A/GAPDH* internal references were calculated using the formula $(2^{-(\text{Ct } QQS - \text{mean Ct internal references})})*100$. Primers are listed in Table S1.

Analysis of DNA methylation by McrBC–qPCR

Total DNA was isolated using Qiagen Plant DNeasy kit following the manufacturer's recommendations. Digestion was carried out overnight at 37°C with 200 ng of genomic DNA and 2 to 8 units of McrBC enzyme (New England Biolabs). Quantitative PCR was performed as described above on equal amounts (2 ng) of digested and undigested DNA samples using

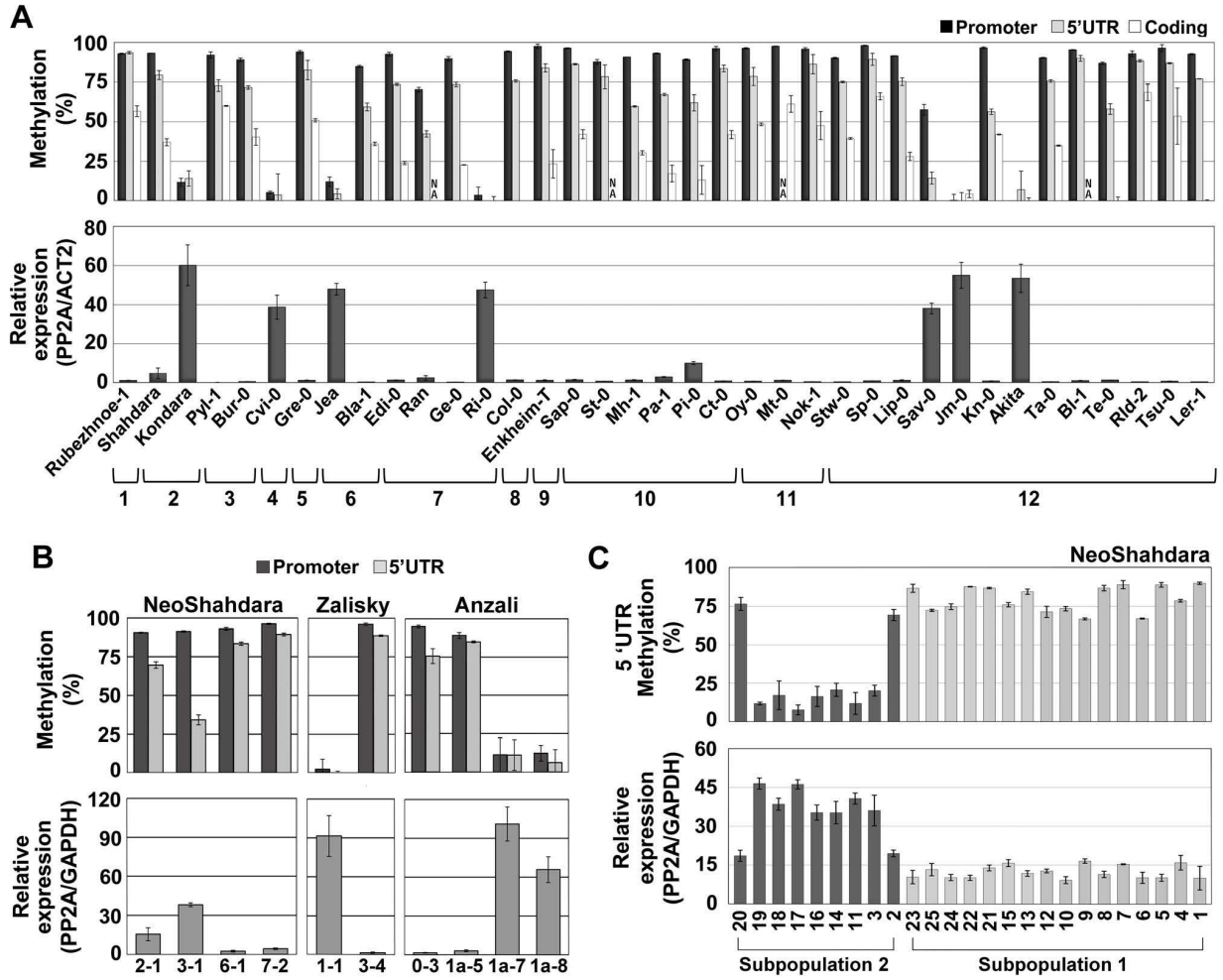


Figure 4. Epigenetic variation at *QQS* is frequent in nature. (A) DNA methylation and expression profile in natural accessions representing the worldwide diversity. Accessions are organized into clades 1 to 12 according to genetic relatedness [36]. NA, not analyzed. (B) DNA methylation and expression levels of *QQS* in plants grown from seeds directly collected in the Central Asian wild populations NeoShahdara, Zalisky and Anzali. For each line, one to three sibling plants were tested and gave similar results so that only one is represented per individual parent. (C) *QQS* epiallelic frequency among 25 NeoShahdara individuals. Plants analyzed here were obtained from seeds produced after two single seed descent generations. Error bars represent standard deviation observed between two (A – DNA methylation) or three (A – expression) biological replicates or two technical replicates (B and C).
doi:10.1371/journal.pgen.1003437.g004

the primers described in Table S1. Results were expressed as percentage of molecules lost through McrBC digestion $(1-(2^{-(\text{Ct digested sample - Ct undigested sample})}))*100$. As a control, the percentage of DNA methylation for *At5g13440*, which is unmethylated in wt, was estimated in all analyses.

Allele-specific expression assays

To assess the relative contribution of each allele to the population of mRNA in F1 individuals from reciprocal crosses between Col and Kondara, a single pyrosequencing reaction using the primers QQS_pyro_F1 (PCR) - TCAAAATGAGGGTCATATC ATGG, QQS_pyro_R1-biotin (PCR) - ATTGGATACAATGGCCCTATAACT and QQS_pyro_S1 (Pyrosequencing) - GATATTGGGCCTTATCAC was set up on a SNP polymorphic between the *QQS* parental coding sequences (Figure S4C; position +285). Pyrosequencing was performed on F1 cDNA, as well as on 1/1 pools of parents cDNA to establish the allelic contribution to *QQS* expression. F1 genomic DNA is used as pyrosequencing control to normalize against possible pyrosequencing biases. Anything significantly driving allele-specific expression in hybrids is by definition acting in *cis*, since F1 nuclei contain a mix of all *trans*-acting factors [43,44].

Comparative genome hybridization (CGH)

CGH experiments were performed for Col-0* vs. Col-0 using Arabidopsis whole-genome NimbleGen tiling arrays [45]. The normalmixEM function of the mixtools package on R was used to found the normal distribution for the distribution of the Col-0*/Col-0 ratio with an expected number of gaussians of two. A Hidden Markov model [46] was used to find regions with copy number variation.

Analysis of genome wide DNA methylation (MeDIP–Chip)

DNA was extracted using DNeasy Qiagen kit and MeDIP-chip was performed on 1.8 µg of DNA as previously described in [47].

The methylated tiles were identified using the ChIPmix method [48]. Probes methylated in one line only (Col-0 or Col-0*) were used to create domains. Domains contain at least three consecutive or nearly consecutive (400 nt min, with one gap of 200 nt max) tiles with identical methylation patterns.

Overall codon-based Z-test of purifying selection

Available *QQS* coding-sequences (464 different accessions) were downloaded from the "Salk Arabidopsis 1001 Genomes" database (http://signal.salk.edu/atg1001/index.php). *A. suecica QQS* sequence (coming from the *A. thaliana* genome of this allotetraploid [49]) was also included in the analysis. The aligned sequences were used to calculate the probability of rejecting the null hypothesis (H_0) of strict-neutrality (dN = dS; where dN = number of non-synonymous and dS = number of synonymous substitutions per site) in favor of the alternative hypothesis of purifying selection (H_A; dS>dN). The analysis was done using the MEGA5 software under the Nei-Gojobori method [50] with the variance of the difference calculated by the bootstrap method with 100 replicates. Our overall analysis of 465 sequences rejected H_0 in favor of H_A (dN/dS = 0.5868; p-value<0.045).

Supporting Information

Figure S1 *QQS* expression is negatively correlated with DNA methylation. (A) Schematic representation of the tandem repeats present at the *QQS* promoter and 5′UTR region. (B) Distribution of DNA methylation at the *QQS* promoter and 5′UTR sequences. Data is presented as the total number of unmethylated (C) and methylated cytosines (5 mC) in the three sequence contexts (CG, CHG and CHH, H = A, T or C) for both DNA strands. DNA methylation data are from [25]. (C) Assessment of *QQS* DNA methylation level and transcript accumulation in seedlings of *ddm1-2* and *rdr2-1* mutants. Error bars represent standard deviation between two (DNA methylation) or three (expression) biological replicates.
(TIF)

Figure S2 Genome-wide analyses of Col-0 and Col-0*. (A) Comparative genomic hybridization (CGH) analysis of Col-0* vs. Col-0 represented as the average of the log 2 ratio of the signal for the INPUT Col-0* over INPUT Col-0. A single normal distribution is observed using the normalmixEM function of the mixtools package on R with an expected number of Gaussians of two. The CGH analyses of Col-0* and Col-0 show no decrease or increase in copy number in Col-0*, suggesting that they correspond to the same accession. In contrast, CGH of Col-0 vs. Cvi and Col-0 vs. C24 revealed 6.0 and 5.5% of tiles with significant copy number variation, respectively [Moghaddam, et al (2011)]. (B–D) Methylated DNA Imunoprecipitation assays. Representation of the proportion of domains that are methylated (B) in only one replicate of Col-0 or in both, (C) in only one replicate of Col-0* or in both and (D) in only Col-0 or Col-0* or in both. A total of 86% of the domains are methylated in both Col-0 and Col-0*, which is similar to the result obtained for two biological replicates of Col-0 or of Col-0* (89% and 91% of the domains methylated in the two replicates, respectively). These results indicate that the methylomes of Col-0* and Col-0 are only marginally more dissimilar from each other as they are from their biological replicates. [Moghaddam A.B, Roudier F, Seifert M, Berard C, Magniette MLM, et al. (2011) Additive inheritance of histone modifications in *Arabidopsis thaliana* intra-specific hybrids. Plant J 67: 691–700. doi: 10.1111/j.1365-313X.2011.04628.x].
(TIF)

Figure S3 Schematic representation of the experimental design used to analyze *QQS* expression and DNA methylation state in single seed descent lines (named Line A, Line B, Line C and so on) at the S1 and S2 generations.
(TIF)

Figure S4 DNA methylation levels of *QQS* correlate negatively with *QQS* expression in natural accessions. (A) Schematic representation of a 30 kb genomic region encompassing *QQS*. Red arrows indicate McrBC-qPCR primer pairs used to determine DNA methylation levels of TEs flanking *QQS*; A, B and C represent different primer pairs designed for the same element. (B) DNA methylation levels of TEs flanking *QQS* in Col-0 (methylated *QQS* epiallele), Jea, Ri-0, Sav-0, Cvi-0, Kondara, Jm-0 and Akita (hypomethylated *QQS* epiallele) accessions. 'NA': not analyzed; 'ND': not determined (presumably because of DNA sequence polymorphisms preventing primer annealing). Error bars represent standard deviation observed in two technical replicates. (C) DNA sequence polymorphisms at the *QQS* locus and flanking region in accessions carrying methylated and hypomethylated *QQS* epialleles. The region analyzed comprises 1.5 kb upstream and 0.6 kb downstream of the *QQS* transcription initiation and termination sites, respectively. Nucleotide positions are numbered relative to the *QQS* translation initiation site (Position +1). Methylated accessions (Col-0, Pyl-1, Mh-1 and Shahdara) are shown in red and hypomethylated accessions (Kondara, Cvi-0, Jea, Ri-0, Sav-0, Jm-0 and Akita) in black. (D) Effect of the methylation inhibitor 5-aza-dC on DNA methylation and expression of *QQS*. Error bars at represent standard deviation observed in at least 3 biological replicates. (E) Pyrosequencing quantification of allele-specific expression of *QQS* in F1 seedlings derived from a cross between Col-0 and Kondara and grown with or without 5-aza-dC. Data is expressed as the % of total transcripts originating from the Kondara allele (top panel). DNA methylation level in the same two pools of F1 seedlings (bottom panel). Error bars represent standard deviation observed in two technical replicates.
(TIF)

Table S1 Primer list.
(DOCX)

Acknowledgments

We are grateful to F. K. Teixeira and members of the V. Colot group for valuable assistance, insights, and discussions. We thank M. Canut for help with the ASE assays.

Author Contributions

Conceived and designed the experiments: ABS CT OL VC MV. Performed the experiments: ABS CT SC JB LEVDB. Analyzed the data: ABS CT SC OL VC MV. Wrote the paper: ABS VC MV.

References

1. Richards EJ (2006) Inherited epigenetic variation - revisiting soft inheritance. Nat Rev Genet 7: 395–401. doi:10.1038/nrg1834.
2. Daxinger L, Whitelaw E (2012) Understanding transgenerational epigenetic inheritance via the gametes in mammals. Nat Rev Genet 13: 153–162. doi:10.1038/nrg3188.
3. Weigel D, Colot V (2012) Epialleles in plant evolution. Genome Biol 13: 249. doi:10.1186/gb-2012-13-10-249.
4. Lippman Z, Gendrel A-V, Black M, Vaughn MW, Dedhia N, et al. (2004) Role of transposable elements in heterochromatin and epigenetic control. Nature 430: 471–476. doi:10.1038/nature02724.1.

5. Kinoshita Y, Saze H, Kinoshita T, Miura A, Soppe WJJ, et al. (2007) Control of *FWA* gene silencing in *Arabidopsis thaliana* by SINE-related direct repeats. Plant J 49: 38–45. doi:10.1111/j.1365-313X.2006.02936.x.
6. Henderson IR, Jacobsen SE (2008) Tandem repeats upstream of the Arabidopsis endogene *SDC* recruit non-CG DNA methylation and initiate siRNA spreading. Genes Dev 22: 1597–1606. doi:10.1101/gad.1667808.
7. Bender J (2004) DNA methylation of the endogenous *PAI* genes in Arabidopsis. Cold Spring Harb Symp Quant Biol 69: 145–153. doi:10.1101/sqb.2004.69.145.
8. Durand S, Bouché N, Strand EP, Loudet O, Camilleri C (2012) Rapid establishment of genetic incompatibility through natural epigenetic variation. Curr Biol 22: 326–331. doi:10.1016/j.cub.2011.12.054.
9. Martin A, Troadec C, Boualem A, Rajab M, Fernandez R, et al. (2009) A transposon-induced epigenetic change leads to sex determination in melon. Nature 461: 1135–1138. doi:10.1038/nature08498.
10. Paszkowski J, Grossniklaus U (2011) Selected aspects of transgenerational epigenetic inheritance and resetting in plants. Curr Opin Plant Biol 14: 195–203. doi:10.1016/j.pbi.2011.01.002.
11. Vaughn MW, Tanurdzić M, Lippman Z, Jiang H, Carrasquillo R, et al. (2007) Epigenetic natural variation in *Arabidopsis thaliana*. PLoS Biol 5: e174. doi:10.1371/journal.pbio.0050174.
12. Zhang X, Shiu S, Cal A, Borevitz JO (2008) Global analysis of genetic, epigenetic and transcriptional polymorphisms in *Arabidopsis thaliana* using whole genome tiling arrays. PLoS Genet 4: e1000032. doi:10.1371/journal.pgen.1000032.
13. Becker C, Hagmann J, Müller J, Koenig D, Stegle O, et al. (2011) Spontaneous epigenetic variation in the *Arabidopsis thaliana* methylome. Nature 480: 245–249. doi:10.1038/nature10555.
14. Schmitz RJ, Schultz MD, Lewsey MG, O'Malley RC, Urich MA, et al. (2011) Transgenerational epigenetic instability is a source of novel methylation variants. Science 334: 369–373. doi:10.1126/science.1212959.
15. Cubas P, Vincent C, Coen E (1999) An epigenetic mutation responsible for natural variation in floral symmetry. Nature 401: 157–161. doi:10.1038/43657
16. Manning K, Tör M, Poole M, Hong Y, Thompson AJ, et al. (2006) A naturally occurring epigenetic mutation in a gene encoding an SBP-box transcription factor inhibits tomato fruit ripening. Nat Genet 38: 948–952. doi:10.1038/ng1841.
17. Miura K, Agetsuma M, Kitano H, Yoshimura A, Matsuoka M, et al. (2009) A metastable *DWARF1* epigenetic mutant affecting plant stature in rice. Proc Natl Acad Sci USA 106: 11218–11223. doi: 10.1073/pnas.0901942106.
18. Donoghue MTA, Keshavaiah C, Swamidatta SH, Spillane C (2011) Evolutionary origins of Brassicaceae specific genes in *Arabidopsis thaliana*. BMC Evol Biol 11: 47. doi:10.1186/1471-2148-11-47.
19. Li L, Foster C, Gan Q, Nettleton D, James MG, et al. (2009) Identification of the novel protein QQS as a component of the starch metabolic network in Arabidopsis leaves. Plant J 58: 485–498. doi:10.1111/j.1365-313X.2009.03793.x.
20. Seo PJ, Kim MJ, Ryu J-Y, Jeong E-Y, Park C-M (2011) Two splice variants of the *IDD14* transcription factor competitively form nonfunctional heterodimers which may regulate starch metabolism. Nat Commun 2: 303. doi:10.1038/ncomms1303.
21. Kaessmann H (2010) Origins, evolution, and phenotypic impact of new genes. Genome Res 20: 1313–1326. doi:10.1101/gr.101386.109.
22. Tautz D, Domazet-Lošo T (2011) The evolutionary origin of orphan genes. Nat Rev Genet 12: 692–702. doi:10.1038/nrg3053.
23. Carvunis A-R, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, et al. (2012) Proto-genes and *de novo* gene birth. Nature 487: 370–374. doi:10.1038/nature11184.
24. Kurihara Y, Matsui A, Kawashima M, Kaminuma E, Ishida J, et al. (2008) Identification of the candidate genes regulated by RNA-directed DNA methylation in Arabidopsis. Biochem Biophys Res Commun 376: 553–557. doi:10.1016/j.bbrc.2008.09.046.
25. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, et al. (2008) Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell 133: 523–536. doi:10.1016/j.cell.2008.03.029.
26. Johannes F, Porcher E, Teixeira FK, Saliba-Colombani V, Simon M, et al. (2009) Assessing the impact of transgenerational epigenetic variation on complex traits. PLoS Genet 5: e1000530. doi:10.1371/journal.pgen.1000530.
27. Colomé-Tatché M, Cortijo S, Wardenaar R, Morgado L, Lahouze B, et al. (2012) Features of the Arabidopsis recombination landscape resulting from the combined loss of sequence variation and DNA methylation. Proc Natl Acad Sci USA 109: 16240–1625. doi:10.1073/pnas.1212955109.
28. Teixeira FK, Heredia F, Sarazin A, Roudier F, Boccara M, et al. (2009) A role for RNAi in the selective correction of DNA methylation defects. Science 323: 1600–1604. doi: 10.1126/science.1165313.
29. Teixeira FK, Colot V (2010) Repeat elements and the Arabidopsis DNA methylation landscape. Heredity 105: 14–23. doi:10.1038/hdy.2010.52.
30. McKhann HI, Camilleri C, Bérard A, Bataillon T, David JL, et al. (2004) Nested core collections maximizing genetic diversity in *Arabidopsis thaliana*. Plant J 38: 193–202. doi:10.1111/j.1365-313X.2004.02034.x.
31. Cubillos FA, Yansouni J, Khalili H, Balzergue S, Elftieh S, et al. (2012) Expression variation in connected recombinant populations of *Arabidopsis thaliana* highlights distinct transcriptome architectures. BMC Genomics 13: 117. doi:10.1186/1471-2164-13-117.
32. Jacobsen SE, Meyerowitz EM (1997) Hypermethylated *SUPERMAN* epigenetic alleles in Arabidopsis. Science 277: 1100–1103. doi:10.1126/science.277.5329.1100.
33. Reinders J, Wulff BBH, Mirouze M, Marí-Ordóñez A, Dapp M, et al. (2009) Compromised stability of DNA methylation and transposon immobilization in mosaic Arabidopsis epigenomes. Genes Dev 23: 939–950. doi:10.1101/gad.524609.
34. Fujimoto R, Kinoshita Y, Kawabe A, Kinoshita T, Takashima K, et al. (2008) Evolution and control of imprinted *FWA* genes in the genus Arabidopsis. PLoS Genet 4: e1000048. doi:10.1371/journal.pgen.1000048.
35. Fujimoto R, Sasaki T, Kudoh H, Taylor JM, Kakutani T, et al. (2011) Epigenetic variation in the *FWA* gene within the genus Arabidopsis. Plant J 66: 831–843. doi:10.1111/j.1365-313X.2011.04549.x.
36. Simon M, Simon A, Martins F, Botran L, Tisné S, et al. (2012) DNA fingerprinting and new tools for fine-scale discrimination of *Arabidopsis thaliana* accessions. Plant J 69: 1094–1101. doi:10.1111/j.1365-313X.2011.04852.x.
37. Kronholm I, Loudet O, Meaux JD (2010) Influence of mutation rate on estimators of genetic differentiation - lessons from *Arabidopsis thaliana*. BMC Genet 11: 33. doi: 10.1186/1471-2156-11-33.
38. Kleinboelting N, Huep G, Kloetgen A, Viehoever P, Weisshaar B (2012) GABI-Kat Simple Search: new features of the *Arabidopsis thaliana* T-DNA mutant database. Nucleic Acids Res 40: D1211–D1215. doi:10.1093/nar/gkr1047.
39. Woody ST, Austin-Phillips S, Amasino RM, Krysan PJ (2007) The WiscDsLox T-DNA collection: an Arabidopsis community resource generated by using an improved high-throughput T-DNA sequencing pipeline. J Plant Res 120: 157–165. doi:10.1007/s10265-006-0048-x.
40. Vongs A, Kakutani T, Martienssen RA, Richards EJ (1993) *Arabidopsis thaliana* DNA methylation mutants. Science 260: 1926–1928.
41. Xie Z, Johansen LK, Gustafson AM, Kasschau KD, Lellis AD, et al. (2004) Genetic and functional diversification of small RNA pathways in plants. PLoS Biol 2: E104. doi:10.1371/journal.pbio.0020104.
42. Oñate-Sánchez L, Vicente-Carbajosa J (2008) DNA-free RNA isolation protocols for *Arabidopsis thaliana*, including seeds and siliques. BMC Res Notes 1: 93. doi:10.1186/1756-0500-1-93.
43. Wittkopp PJ, Haerum BK, Clark AG (2004) Evolutionary changes in *cis* and *trans* gene regulation. Nature 430: 85–88. doi:10.1038/nature02698.
44. Zhang X, Richards EJ, Borevitz JO (2007) Genetic and epigenetic dissection of *cis* regulatory variation. Curr Opin Plant Biol 10: 142–148. doi:10.1016/j.pbi.2007.02.002.
45. Moghaddam A.B, Roudier F, Seifert M, Berard C, Magniette MLM, et al. (2011) Additive inheritance of histone modifications in *Arabidopsis thaliana* intra-specific hybrids. Plant J 67: 691–700. doi: 10.1111/j.1365-313X.2011.04628.x
46. Seifert M, Banaei A, Grosse I, Stricken M (2009) Array-based comparison of Arabidopsis ecotypes using hidden Markov models. In: Encarnação P, Veloso A, editors. BIOSIGNALS 2009. Portugal: INSTICC Press. pp. 3–11.
47. Cortijo S, Wardenaar R, Colome-Tatche M, Johannes F, Colot V (2012) Genome-wide analysis of DNA methylation in Arabidopsis using MeDIP-chip. In: McKeown PC and Spillane C, editors. Treasuring Exceptions: Plant Epigenetics and Epigenomics. New Jersey: Humana Press. *In press*
48. Martin-Magniette M L, Mary-Huard T, Berard C, Robin C (2008) ChIPmix: mixture model of regressions for two-color ChIPchip analysis. Bioinformatics 24: I181–I186. doi:10.1093/bioinformatics/btn280.
49. Jakobsson M, Hagenblad J, Tavaré S, Säll T, Halldén C, Lind-Halldén C, and Nordborg M (2006). A unique recent origin of the allotetraploid species *Arabidopsis suecica*: Evidence from nuclear DNA markers. Mol Biol Evol 23: 1217–1231. doi:10.1093/molbev/msk006.
50. Nei M, Gojoborit T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3: 418–426.
51. Buisine N, Quenesville H, Colot V (2008) Improved detection and annotation of transposable elements in sequenced genomes using multiple reference sequence sets. Genomics 91: 467–475. doi: 10.1016/j.ygeno.2008.01.005.
52. Ahmed I, Sarazin A, Bowler C, Colot V, Quenesville H (2011) Genome-wide evidence for local DNA methylation spreading from small RNA-targeted sequences in Arabidopsis. Nucleic Acids Res 39: 1–13. doi: 10.1093/nar/gkr324.

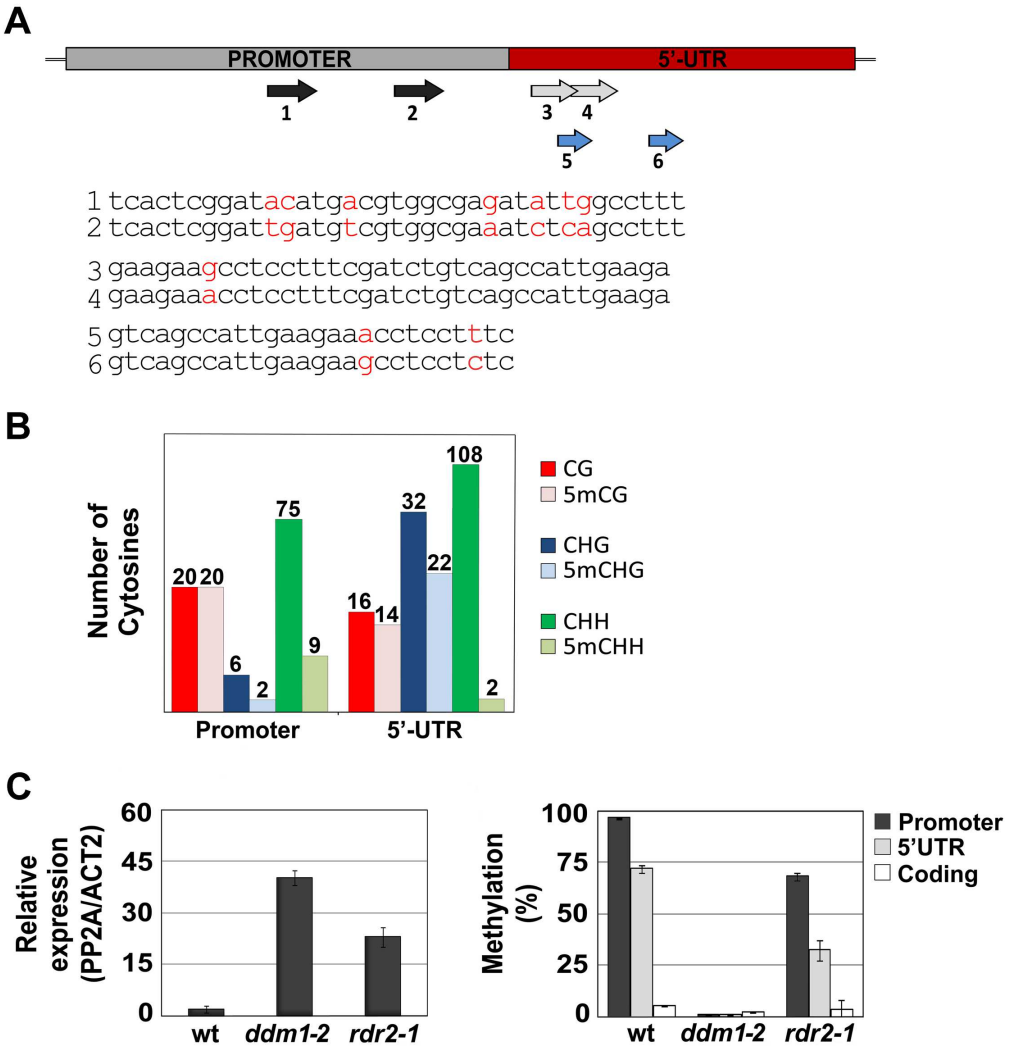


FIG.S1. ***QQS* expression is negatively correlated with DNA methylation.** (A) Schematic representation of the tandem repeats present at the *QQS* promoter and 5-UTR region. (B) Distribution of DNA methylation at the *QQS* promoter and 5-UTR sequences. Data is presented as the total number of unmethylated (C) and methylated cytosines (5mC) in the three sequence contexts (CG, CHG and CHH, H = A, T or C) for both DNA strands. DNA methylation data are from [25]. (C) Assessment of *QQS* DNA methylation level and transcript accumulation in seedlings of *ddm1-2* and *rdr2-1* mutants. Error bars represent standard deviation between two (methylation) or three (expression) biological replicates.

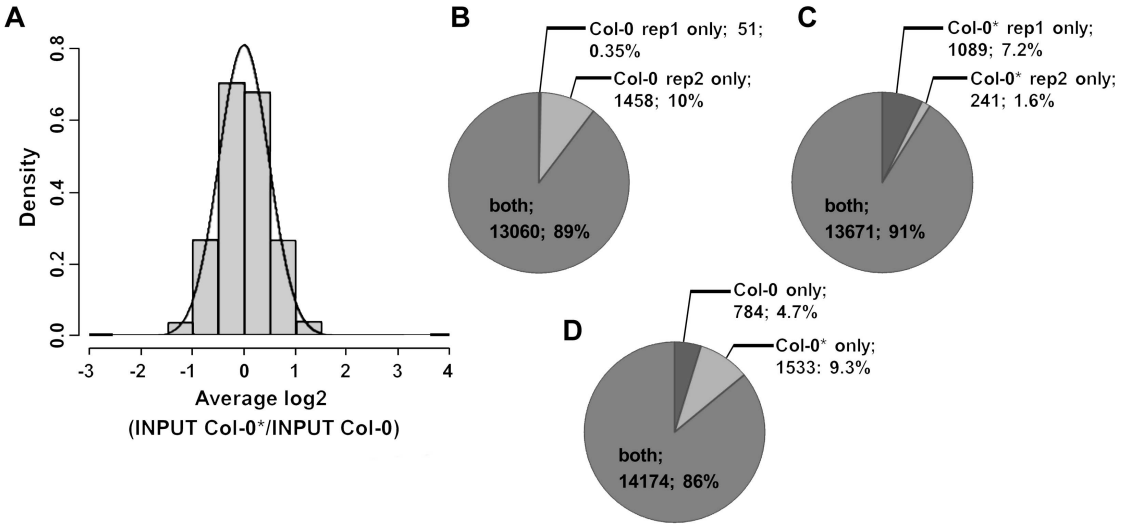


FIG.S2. **Genome-wide analysis of Col-0 and Col-0*.** (A) Comparative genomic hybridization (CGH) analysis of Col-0* vs. Col-0 represented as the average of the log 2 ratio of the signal for the INPUT Col-0* over INPUT Col-0. A single normal distribution is observed using the normalmixEM function of the mixtools package on R with an expected number of Gaussians of two. The CGH analyses of Col-0* and Col-0 show no decrease or increase in copy number in Col-0*, suggesting that they correspond to the same accession. In contrast, CGH of Col-0 vs. Cvi and Col-0 vs. C24 revealed 6.0 and 5.5% of tiles with significant copy number variation, respectively [Moghaddam, et al (2011)]. (BD) Methylated DNA Imunoprecipitation assays. Representation of the proportion of domains that are methylated (B) in only one replicate of Col-0 or in both, (C) in only one replicate of Col-0* or in both and (D) in only Col-0 or in Col-0* or in both. A total of 86% of the domains are methylated in both Col-0 and Col-0*, which is similar to the result obtained for two biological replicates of Col-0 or of Col-0* (89% and 91% of the domains methylated in the two replicates, respectively). These results indicate that the methylomes of Col-0* and Col-0 are only marginally more dissimilar from each other as they are from their biological replicates.

Moghaddam A.B, Roudier F, Seifert M, Berard C, Magniette MLM, et al. (2011) Additive inheritance of histone modifications in *Arabidopsis thaliana* intra-specific hybrids. Plant J 67: 691-700. doi: 10.1111/j.1365-313X.2011.04628.x

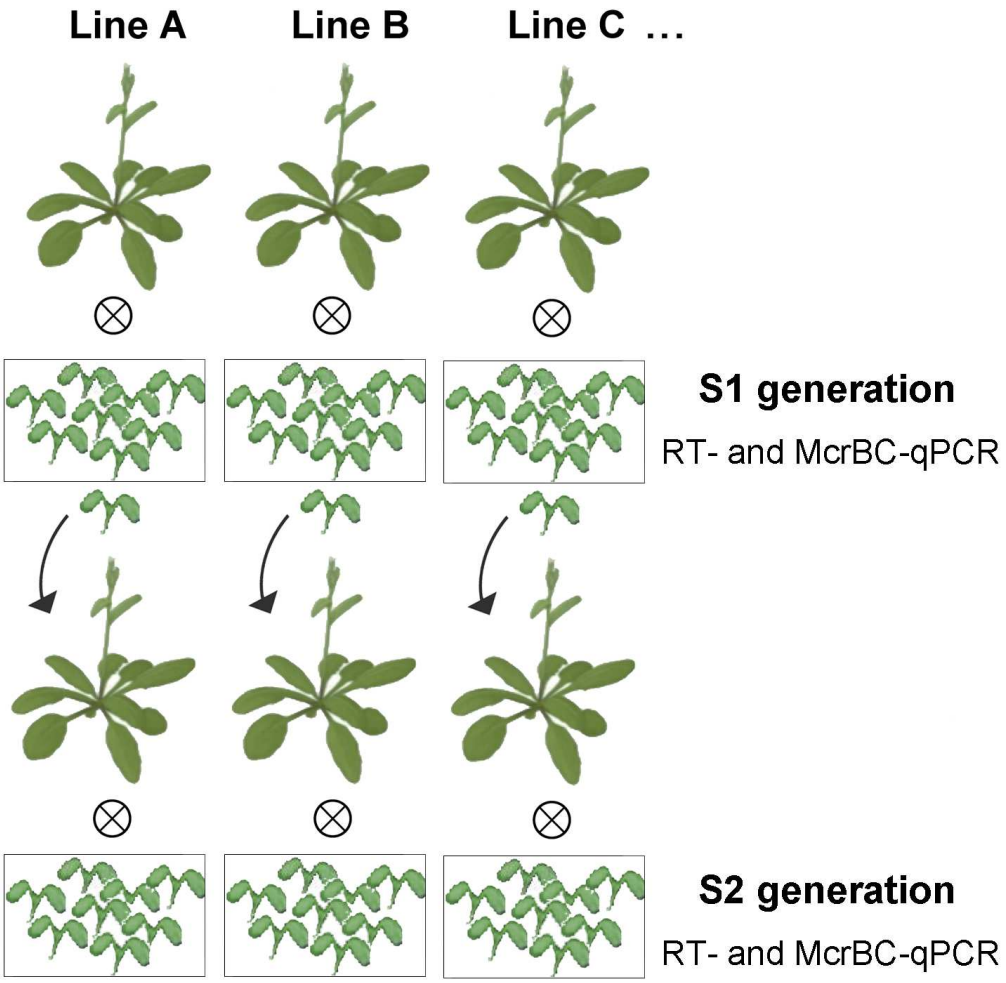


FIG.S3. Schematic representation of the experimental design used to analyze *QQS* expression and DNA methylation state in single seed descent lines (named Line A, Line B, Line C and so on) at S1 and S2 generations.

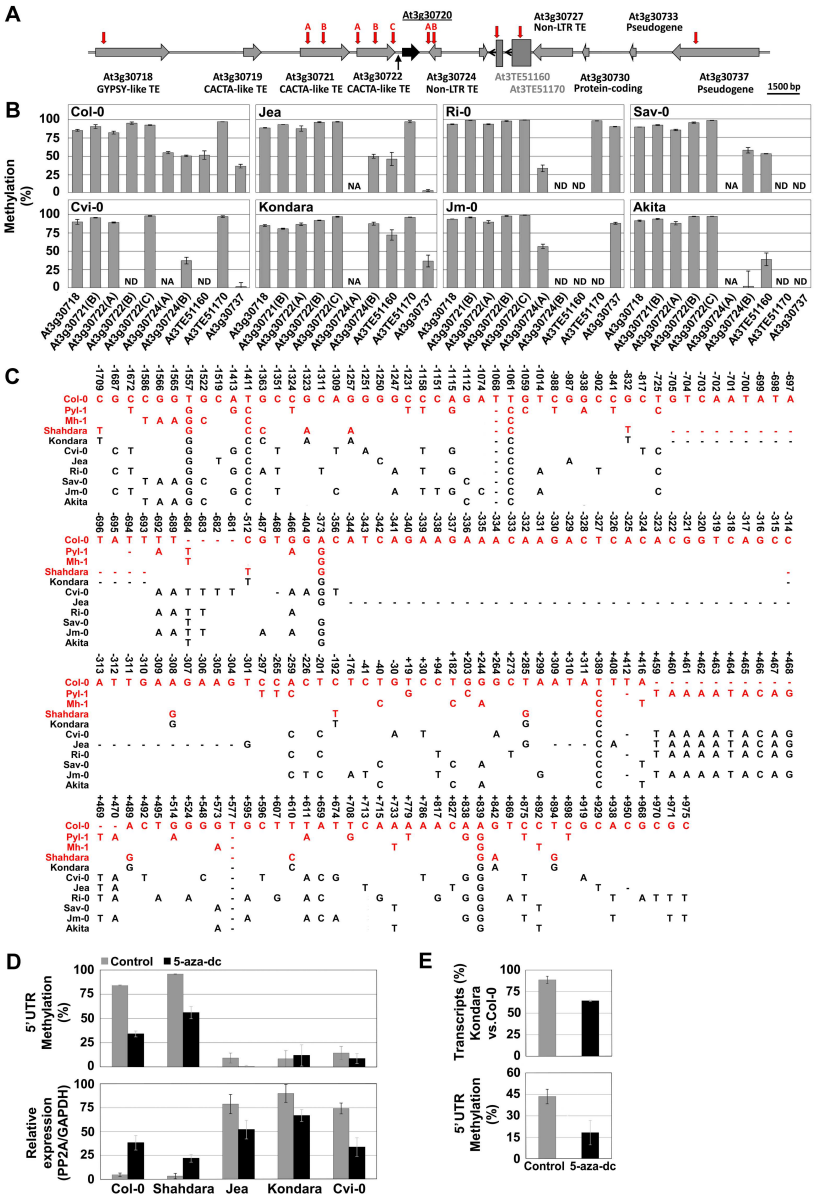


FIG.S4. **DNA methylation levels of *QQS* correlate negatively with *QQS* expression in natural accessions.** (A) Schematic representation of a 30 kb genomic region encompassing *QQS*. Red arrows indicate McrBC-qPCR primer pairs used to determine DNA methylation levels of TEs flanking *QQS*; A, B and C represent different primer pairs designed for the same element. (B) DNA methylation levels of TEs flanking *QQS* in Col-0 (methylated *QQS* epiallele), Jea, Ri-0, Sav-0, Cvi-0, Kondara, Jm-0 and Akita (hypomethylated *QQS* epiallele) accessions. NA: not analyzed; ND: not determined (presumably because of DNA sequence polymorphisms preventing primer annealing). Error bars represent standard deviation observed in two technical replicates. (C) DNA sequence polymorphisms at the *QQS* locus and flanking region in accessions carrying methylated and hypomethylated *QQS* epialleles. The region analyzed comprises 1.5 kb upstream and 0.6 kb downstream of the *QQS* transcription initiation and termination sites, respectively. Nucleotide positions are numbered relative to the *QQS* translation initiation site (Position +1). Methylated accessions (Col-0, Pyl-1, Mh-1 and Shahdara) are shown in red and hypomethylated accessions (Kondara, Cvi-0, Jea, Ri-0, Sav-0, Jm-0 and Akita) in black. (D) Effect of the methylation inhibitor 5-aza-dC on DNA methylation and expression of *QQS*. Error bars at represent standard deviation observed in at least 3 biological replicates. (E) Pyrosequencing quantification of allele-specific expression of QQS in F1 seedlings derived from a cross between Col-0 and Kondara and grown with or without 5-aza-dC. Data is expressed as the % of total transcripts originating from the Kondara allele (top panel). DNA methylation level in the same two pools of F1 seedlings (bottom panel). Error bars represent standard deviation observed in two technical replicates.

Expression analysis				
AGI	**Primer pair**	**Forward primer**	**Reverse primer**	**Experiments**
At3g30720 (QQS)	QQS_1	AAGACCAATAGAGAGCAGGAA	CCTGATGTAGAAGTGTGAGG	All, with the exception of WiscDsLoxHs077_09G genotype on Figure 3
	QQS_2	CACTTCTACATCAGGTGTCG	AAGGCCCAATATCAGTAGTTG	WiscDsLoxHs077_09G on Figure 3
AT1G13320 (PP2A)	PP2A_1	CATGTTCCAAACTCTTACCTG	GTTCTCCACAACCGCTTGGT	Figures 2, 3, 4A, S1
	PP2A_2	TTTGTGAAGCTGTAGGACCG	CGAGTTCAGGGTTTAAAATGCG	Figures 4B, 4C and S4
AT3G18780 (Actin2)	ACT2	GTACAACCGGTATTGTGCTGG	CAAGGTCAAGACGGAGGATG	Figures 2, 3, 4A, S1
At1g13440(GAPDH)	GAPDH	TTGGTGACAACAGGTCAAGC	AAACTTGTCGCTCAATGCAATC	Figures 4B, 4C and S4
DNA methylation quantification				
AGI	**Primer pair**	**Forward primer**	**Reverse primer**	**Experiments**
At3g30720(QQS)	Promoter (F1A/R1A)	TCACTCGGATTGATGTCGTG	AGGAGACGAAACAGACAAATC	Figures 2, 3 (with the exception of GABI-Kat_522C07), 4 (Col-0, Gre-0, Ct-1; Enkheim-0, Jea, Ler-1, Nok-1, Pa-1, Pi-0, Sp-0, Bur-0, Edi-0, Te-0, Mh-1, Sav-0, Mt-0, Akita, Kondara, Shahdara, Tsu-0, NeoShahdara, Zalisky and Anzali) and S1
	Promoter (F1A/R1B)	TCACTCGGATTGATGTCGTG	AGGAGATGAAACAGACAAATC	Figure 4 (Cvi-0, Bl-1, Bla-1, Ge-0, Pyl-1, Oy-0, Jm-0, Kn-0, Lip-0, Rubezhoe, Sap-0, Ta-0, Rld-2, Stw-0)
	Promoter (F1B/R1B)	GCCAATTAGAATGTTTCACTCG	AGGAGATGAAACAGACAAATC	Figure 4 (Ran, St-0)
	Promoter (F1C/R1A)	TCCAAGCTTGCCAAAACGATC	AGGAGACGAAACAGACAAATC	Figure 3 (GABI-Kat_522C07)
	5'-UTR (F2A/R2A)	TCTGTCAGCCATTGAAGAAAC	GATAAGGTTTGGGTACAGATC	Figures 2, 3, 4 (Col-0, Ri-0, Gre-0, Ct-1, Enkheim-T, Ler-1, Nok-1, Pa-1, Pi-0, Ran, Sp-0, Bur-0, St-0, Mh-1, Sav-0, Ta-0, Mt-0, Akita, Kondara, Shahdara, Tsu-0, NeoShahdara, Zalisky and Anzali), S1 and S4
	5'-UTR (F2A/R2B)	TCTGTCAGCCATTGAAGAAAC	GATAAGGGTTGGGTACAGATC	Figure 4 (Cvi-0, Bl-1, Bla-1, Ge-0, Pyl-1, Edi-0, Oy-0, Te-0, Jm-0, Lip-0, Rubezhnoe, Sap-0, Rld-2, Stw-0)
	5'-UTR (F2B/R2B)	TCTGTCAGCCATTGAAGAAGC	GATAAGGGTTGGGTACAGATC	Figure 4 (Kn-0)
	5'-UTR (F2B/R2C)	TCTGTCAGCCATTGAAGAAGC	GATAAGGTTTGGGCACAGATC	Figure 4 (Jea)
	Coding (F3A/R3A)	AAGGTTCATTTTGCCTCACAC	AAGGCCCAATATCAGTAGTTG	Figures 2, 3 (with the exception of WiscDsLoxHs077_09G), 4 (Col-0, Gre-0, Ri-0, Bla-1, Ct-1, Enkheim-T, Ge-0, Jea, Ler-1, Nok-1, Pyl-1, Sp-0, Bur-, Edi-0, Oy-0, Te-0, Jm-0, Kn-0, Lip-0, Rubezhnoe, Sap-0, Ta-0, Mt-0, Kondara, Rld-2, Shahdara, Stw-0, Tsu-0, NeoShahdara, Zalisky and Anzali) and S1
	Coding (F3B/R3A)	AAGGTTCATTCTGCCTCACAC	AAGGCCCAATATCAGTAGTTG	Figure 4 (Pa-1, Pi-0, Mh-1, Sav-0, Cvi-0, Akita)
At5g13440	At5g13440	ACAAGCCAATTTTTGCTGAGC	ACAACAGTCCGAGTGTCATGGT	All
At3g30722	At3g30722(A)	GCCGTAGTAACCGTCAGGAA	AGACATTTTATTCTGTTAAGTGG	As indicated on Figures 3 and S4
	At3g30722(B)	CTGCTAGAATGGGGTTCATC	CCTCCATAGTGGCGAATCAC	
	At3g30722(C)	GTTAGACTACAAGTACCAACTC	AAGAGTTGCAGGATCCGTCG	
At3g30721	At3g30721(A)	CTCTGGAGCATCAATTAGTTTG	ACTTCAAATCCATACCTCTGAT	
	At3g30721(B)	GAGAAACCTTCGTCTTGGTC	GGGATCAACATAGTCAACATG	
At3g30718	At3g30718	GTCTAGATATCCAGGGGATG	CTCTGAACTATCAACATGTGC	
At3g30724	At3g30724(A)	TCCTGTGCTATTGATACTCAC	GACAAAACAAGTCTGATCGATG	
	At3g30724(B)	TCCTGTGCTATTGATACTCAC	GATGTTTTTGCAGTCAATGAAAC	
AT3TE51160	AT3TE51160	CTTTACTTACAAAGTAGATGAGC	CTCGAGATTGACTCTTTTGAG	
At3TE51170	At3TE51170	CCTGTGATATACCGTCTCGT	GGTCGATAGTAATACGAGAGA	
At3g30737	AT3G30737	AGTGTCTCGAACGTGTGTCG	GTTACAGAAGATTCCATTTGTG	

TABLE.S1. Primer list.

13. GENERAL DISCUSSION AND PERSPECTIVES

13.1 The phenotypic consequences of QQS epivariants

An important and outstanding question in our study is the phenotypic consequences of *QQS* epialleles and how those epivariants contribute to transgenerational epigenetic variation. From two previous papers, QQS protein likely contributes to the regulation of starch metabolism. Starch is an important polysaccharide that accumulates linearly in chloroplasts during the light period when photosynthates are abundant to be then linearly degraded during the night. A tight control of starch degradation by the circadian clock allows a continuous reallocation of carbohydrates during the dark period to support metabolism and growth so that all the starch accumulated during the day is almost completely degraded at dawn. Overall, the regulation of starch turnover is essential to maximize photosynthates partitioning and usage during the day and to avoid starvation at night so that plants maintain growth throughout 24h [342, 343]. In *QQS*-targeting RNAi lines (Col-0 background), an excess of 20 to 30% of leaf starch content was observed under long-day conditions at the end of the light period [341], whereas *QQS* overexpressing lines (using CaMV-35s promoter in Col-0 background) presented a reduction of 20% of their starch content in similar conditions [344]. Those results suggested that QQS is a negative regulator of starch accumulation during the light period. As starch content is an important factor controlling biomass accumulation [342], variation in *QQS* expression level could contribute to the regulation of starch turnover and growth, a regulation that could potentially be adaptive. Thus we analysed growth and starch content under various conditions in lines expressing *QQS* at different levels.

First, on a standard *in vitro* media with sucrose, under long day conditions, we could not detect significant shoot growth differences between HIFs fixed Col-0 (lowly expressed and highly methylated *QQS* allele) or Cvi-0 (highly expressed and lowly methylated *QQS* alleles) at *QQS* (figure 40A), nor between epiRILs carrying methylated/unmethylated *QQS* epialleles (figure 40B). Besides, no clear growth difference has been observed between Col-0* lines relative to their *QQS* expression level, neither between a T-DNA mutant likely disrupting QQS protein (WiscHS077_09G) and its WT line (figure 40C-D). Because the sucrose frequently added to the *in vitro* media could interfere with starch metabolism and growth, we also confirmed the previous observations on a media containing no sucrose (figure 40C-D). Under those conditions,

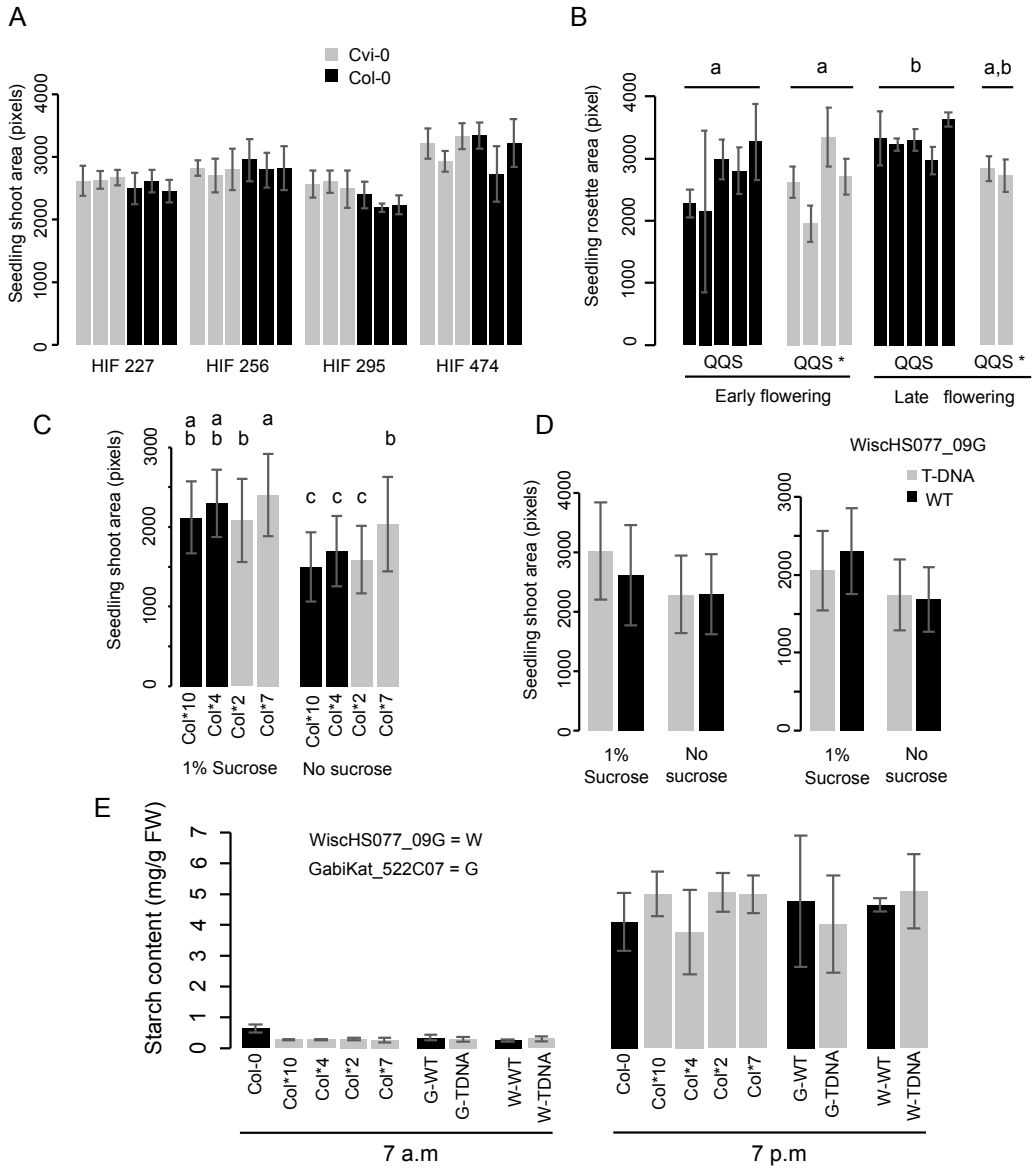


Fig. 40. *QQS* epialleles have no particular phenotype under long-day *in vitro* conditions. HIFs, epiRILs and Col-0* lines were used to compare the shoot area of plants with highly methylated (black) or lowly methylated (grey) *QQS* epialleles (A-C) *in vitro* on standard media with or without 1% sucrose. Different HIFs (A - 3 different individuals of the same genotype are represented), epiRILs (B - each bar corresponds to an epiRIL) and Col-0* lines (C) were phenotyped to analyse the effect of different genetic and epigenetic background on the phenotype. D. The phenotyping of a T-DNA insertion line responsible for reduced *QQS* expression level compared to Col-0 (WiscHS077_09G) revealed no particular phenotype *in vitro* on standard media with or without 1% sucrose. Two biological replicates are shown. Error bars represent standard errors obtained from the phenotyping of at least 30 12 DAS old seedlings (A-D). E. Starch content have been measured in lines with different *QQS* alleles and epialleles at dawn (7 a.m) and after 12 hours of light (7 p.m). GabiKat_522C07 express *QQS* at a higher level than Col-0 due to an insertion in *QQS* promoter. For each individual line, error bars represent standard errors obtained from the phenotyping of 3 pools of 12 DAS seedlings grown in long–day conditions.

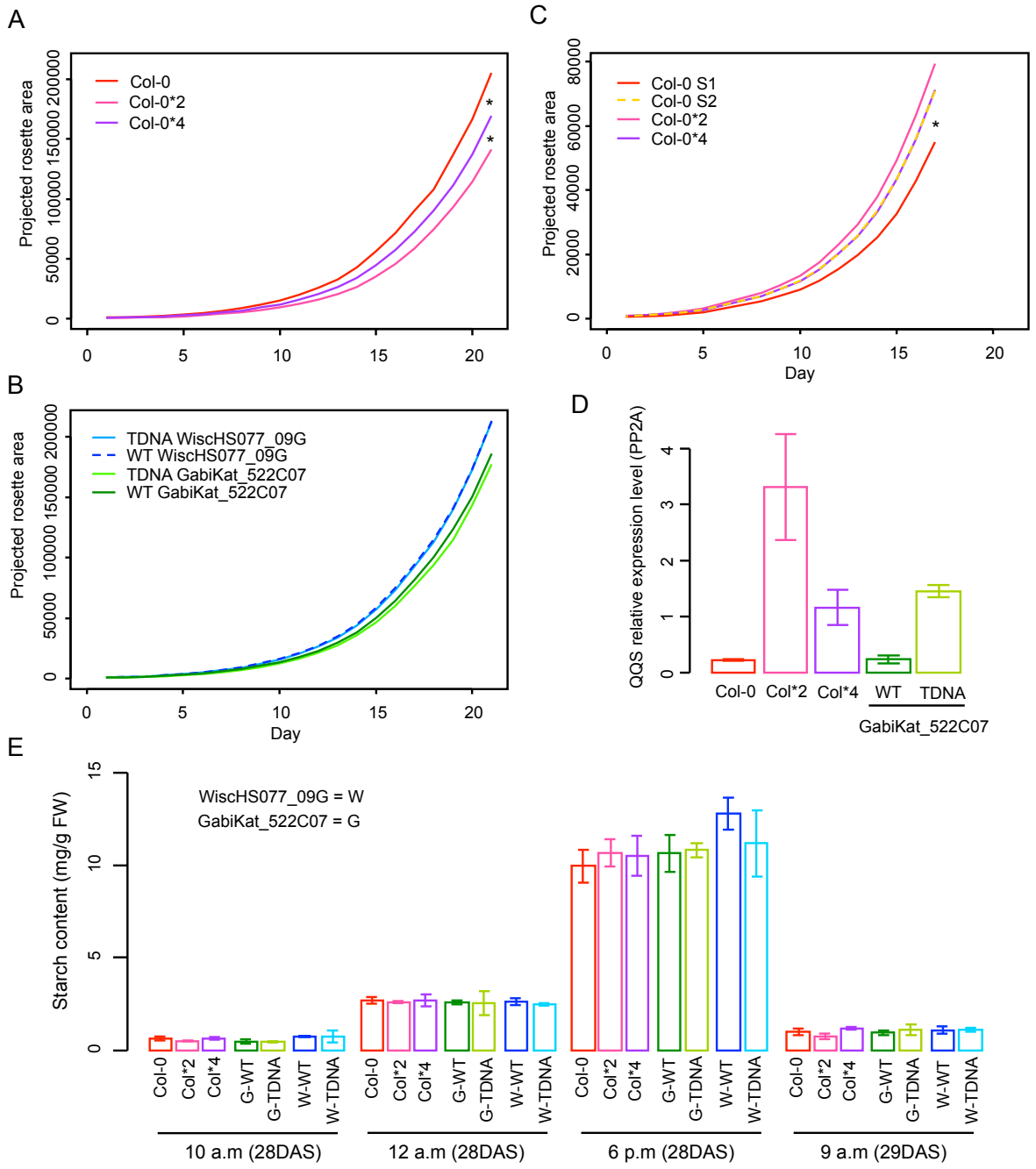


Fig. 41. *QQS* epialleles have no particular phenotype under short-day *in vivo* conditions. A-C. Phenotyping on the PHENOSCOPE in short-day conditions of several lines with different *QQS* alleles and epialleles. The slight difference observed between Col-0, Col-0*4 and Col-0*2 in the first experiment (A, ANOVA $p < 0.05$) was not confirmed in the second one (C). Both T-DNA insertion lines present no phenotype compared to WT (B). Day 0 correspond to the day of installation on the PHENOSCOPE and at that time the plants are already 8 DAS old. D. Relative *QQS* expression levels in 28 DAS plants grown on the phenoscope. *PP2A* endogenous control was used to determine expression levels of *QQS* using the $2^{(Ct_{gene} - Ct_{PP2A})}$. Error bars represent the standard deviation observed in three individual plants. E. Starch content have been measured in several lines with different *QQS* alleles and epialleles grown on the phenoscope in short-days condition (light from 10 a.m to 6 p.m). For each individual line, error bars represent standard errors obtained from the phenotyping of pools of leaves 6, 7 and 8 of 3 independent individuals.

we finally measured starch accumulation in the different Col-0* lines and T-DNA mutants (WiscHS077_09G and another T-DNA line - GabiKat_522C07- whose insertion in the promoter of *QQS* results in increase transcript accumulation) during the course of the day and did not observe a consistent increase or decrease of starch content at the end of the light period as expected regarding previous reported results (figure 40E). We also analysed the projected rosette area of plants grown on real soil on the PHENOSCOPE, a high-throughput phenotyping robot that waters, rotates and takes pictures of the plants automatically every day under tightly controlled short-day conditions [76]. Under those conditions, in a first experiment, we observed that Col-0 plants were slightly bigger than Col-0*4 plants that were also slightly bigger than Col-0*2 plants (Col-0*4 being partially demethylated and Col-0*2 being totally demethylated at *QQS*) (figure 41A) suggesting that when *QQS* is highly expressed, the plant grew slower than when it is downregulated. However, no difference was observed between the two T-DNA mutants and their WT (figure 41B). Besides, those fine results could not be confirmed in a second experiment (figure 41C). To see if starch content differences could be observed in those conditions, we performed starch dosages in shoots of the different lines at different time points during the day. Here again, no difference could be observed (figure 41E).

Overall, the absence of starch phenotype in our Col-0* and T-DNA lines was surprising. Regarding what had been found before, we were expecting to see an increase of starch content in the WiscHS077_09G T-DNA line compared to WT as observed before for *QQS* RNAi lines, and a decrease in starch content in the Col-0* and GabiKat_522C07 T-DNA insertion lines compared to Col-0 and WT as these lines are characterised by an increase in *QQS* expression level (figure 41D) as described previously in *p35s*-overexpressor lines. Those results could be explained by environmental differences that could influence starch accumulation (long-day vs short-day conditions, soil vs *in vitro* conditions or different *in vitro* media). Nevertheless, Amanda tested on her side (data not shown) conditions very similar to the one used in the Seo and colleagues' paper [344] and the various conditions I tested reinforce the idea that the absence of difference in starch accumulation in our lines is not environmental. To properly test the absence of environmental effect, we should have used *QQS* RNAi or pCaMV-35s:*QQS* lines as positive control in our experiments. Conversely, differences between our results and the one of the previous *QQS* studies regarding starch accumulation could be genetic. But then, what distinguishes Col-0* and GabiKat_522C07 T-DNA insertion lines from the *pCaMV-35s:QQS* overexpressors [344] and what are the differences between WiscHS077_09G T-DNA mutant and the RNAi line [341] Compared to *pCaMV-35s:QQS* overexpressor, Col-0* lines and -to a lesser extent- the GabiKat_522C07 T-DNA insertion line, despite their high expression level, could still be regulated by potential trans-factors. This absence of regulation in *pCaMV-35s:QQS* overexpressor and *QQS* RNAi line might be what drives the differential starch accumulation observed in those lines. One potential important factor could be the circadian clock that likely

regulates diurnal patterns of *QQS* transcript accumulation according to the website 'Diurnal' and that is essential for the control of starch degradation at night (although no evidence of its role in starch partitioning during the light period has been obtained yet; [342]). However this hypothesis likely does not explain the absence of phenotype in the WiscHS077_09G T-DNA mutant.

QQS is known to respond transcriptionally to a variety of developmental, genetic and environmental perturbations. Among them, cold treatment was shown to be responsible for *QQS* downregulation and Seo and colleagues suggested that *QQS* regulation could be important to maintain correct levels of starch under cold treatment [344]. Thus we analysed the effect of cold on *QQS* expression level and starch content in lines with different *QQS* alleles and epialleles. We showed that the intensity of *QQS* repression induced by the cold treatment was independent of *QQS* methylation level (figure 42A). This may involve several kinds of mechanisms including mRNA degradation coupled or not with transcriptional repression. We did not observe either any significant difference in starch accumulation between the different *QQS* alleles and epialleles (figure 42B). However, it is important to note that our conditions were likely not ideal to detect QQS effect on starch accumulation as no starch increase has been observed during the light period under the cold treatment as opposed to what was previously reported [344, 345]. This might be the result of low light intensity in our growth chamber at 4 °C.

As a conclusion, our phenotypic analyses suggest that *QQS* epialleles had no impact on starch accumulation or growth. This conclusion is consistent with some QTL mapping experiments that show that the *QQS* region is not associated with any major QTL regarding starch accumulation [91, 346]. The absence of integrative phenotypic effect could indicate that those epivariants are not adaptive under natural environments. Consistently, methylated and unmethylated epialleles at *QQS* (see publication – Fig. 4 and figure 45) have been recurrently observed in several *A. thaliana* local populations. At such small scales (meters), all plants are expected to undergo the same environmental constraints and this is even truer regarding day length, temperatures and light quality which are likely the most important climatic variables that influence starch accumulation. However, it is important to note that *QQS* epialleles might play a significant role in the regulation of starch metabolism in untested specific environmental conditions, organs or developmental windows. For example, the effect of the clock on starch accumulation in Col-0* lines could be analysed more precisely. Similarly, starch was shown to be particularly important regarding the regulation of leaf expansion rate in young leaves compared to mature leaves [347] so that looking specifically at the growth of this organ could have been interesting. Besides, any conclusion regarding adaptation would need to link starch perturbations with growth defects and ultimately fitness variations. Finally, epigenetic variation could actually be maintained in local population as a result of balancing selection over time. Analysing the evolution of the epivariants frequencies over time in several natural or

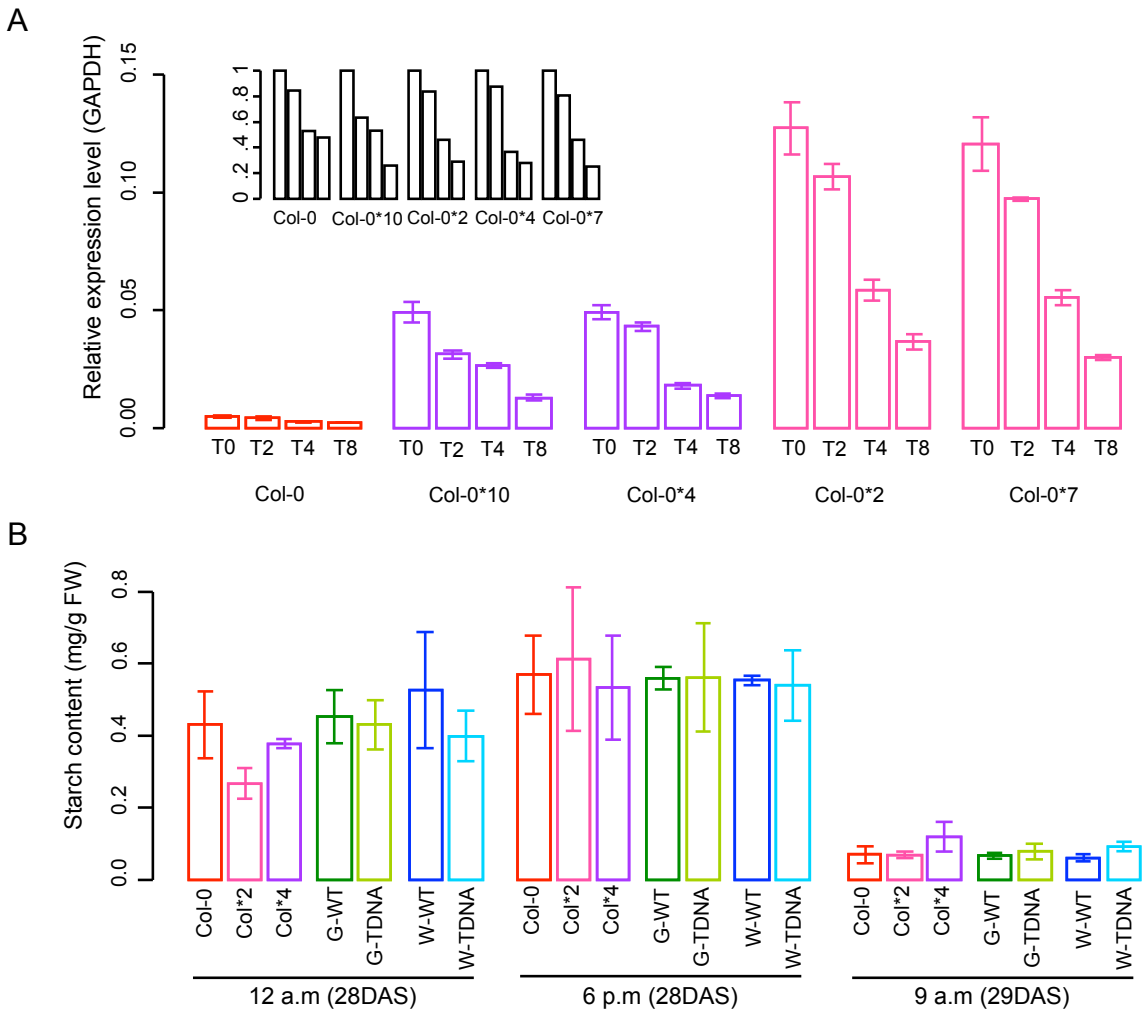


Fig. 42. *QQS* cold response at the transcriptomic and phenotypic level. A. Relative *QQS* expression levels in 28 DAS plants grown in the greenhouse under long-day conditions and transferred at 4 °C in a growth chamber for 2 to 8 hours. *GAPDH* endogenous control was used to determine expression levels of *QQS* using the $2^{(Ct_{gene}-Ct_{GAPDH})}$. Error bars represent the standard deviation observed in two technical replicates. In the upper diagrams the expression level has been reported to the one of T0 to highlight the fact that cold response is similar in the different lines. A single individual was measured for each point. B. Starch content have been measured in different lines in 28 DAS plants grown on the phenoscope in short-day conditions (light from 10 a.m to 6 p.m) and transferred at 4 °C (at 10 a.m) for 2, 8 and 23 hours. For each individual line, error bars represent standard errors obtained from the phenotyping of 3 individual plants.

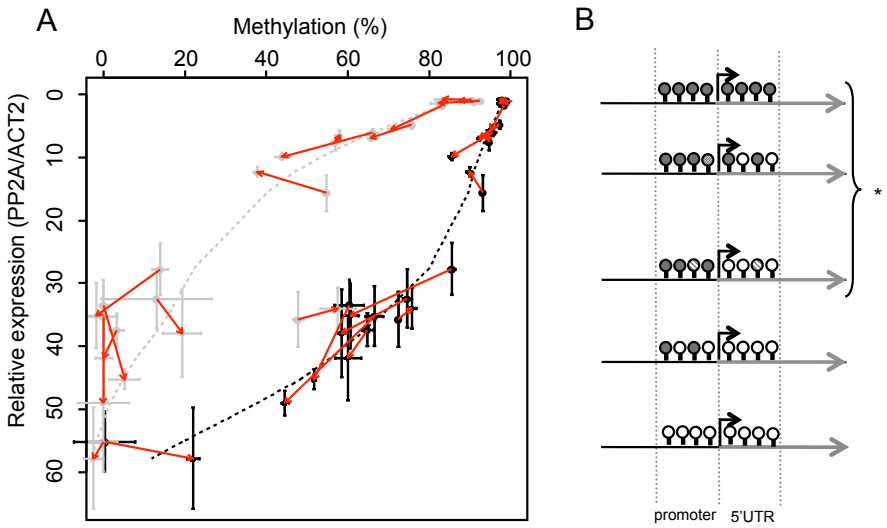


Fig. 43. Col-0* epialleles result from the simultaneous loss of methylation in the 5' UTR and in the promoter of *QQS* gene. Diagram A. sum up the results of the figures 2B and 2C of the PLoS Genetics paper. Red arrows indicate S1 to S2 methylation/expression variation for each Col-0* line. The data for the 5' UTR are indicated in grey and the one for the promoter in black. B. Schemes representing sequential demethylation of 5' UTR and promoter. * Note that compared to the Col-0* bulk, all the NeoSha individuals are methylated for *QQS* promoter and varied solely in 5' UTR methylation level.

experimentally-controlled environments could help deciphering the adaptive potential of *QQS* epialleles as well as their evolution (emergence and maintenance).

13.2 The origin and evolution of QQS epivariants

The occurrence and evolution of *QQS* epialleles remains rather elusive. First, as mentioned in the introduction, identifying the origin of epivariants can be tricky, particularly regarding the role of inducing genetic modifications. In our case, although the origin of *QQS* methylation switches is unknown, several pieces of evidence tend to suggest that *QQS* is likely a pure epiallele. First, our data clearly showed that at instant 't'across evolutionary time, the segregation of *QQS* epivariants in Col-0* lines, RILs and accessions was independent of any genetic variation in the proximity of *QQS* gene (no local sequence variation between Col-0 and Col-0* as well as between Shahdara and Kondara; 5-aza-DC experiments; independent flanking TE methylation status). Besides, data from eQTL mapping and ASE tests clearly pointed out that most of the variation in expression levels between Col-0 and Jea, Cvi-0 and Kondara results from a local cis-effect reinforcing the idea that *QQS* epivariants do not result from the immediate action of a major trans-acting factor. Still, *QQS* epivariants could result from a trans acting genetic mutation that does not segregate anymore in the accessions. However

this hypothesis is unlikely because few mutations are expecting to segregate in the Mutation Accumulation (MA) lines (where variation at *QQS* methylation was detected) and between Col-0 and Col-0* lines [43]. Besides, it would suggest that one or several genetic variants with the same phenotypic consequence (*QQS* promoter methylation/expression) appeared independently in different lineages. Overall, *QQS* epivariants probably arose via incorrect maintenance of DNA methylation possibly via or concomitant with the loss of small RNAs targeting *QQS*, but the exact molecular mechanism remains to be determined.

Our results regarding the stability of *QQS* promoter and 5' UTR methylation are rather contrasted. Indeed, we show that methylated and demethylated epialleles are relatively stable for several generations (eQTL in Col-0 x Cvi-0 RIL population (8 generations), Col-0* lines (2 generations)) but also switch at a non negligible frequency (in MA lines, Col-0 vs Col-0* laboratory seed stocks, Edi-0 accession (methylated in our analysis and likely demethylated in the one of Gan and colleagues (MAGIC gbrowse [41]). Besides, the direction of methylation variations is not clear. All switches we observed were from a methylated status to a demethylated one (MA and Col-0 vs Col-0*) but we don't know if a demethylated allele could get remethylated at some point. This latter possibility makes sense considering the high frequency of methylated alleles in natural accessions as well as the relatively high frequency of *QQS* demethylation events. The identification of 'intermediate' epialleles in Col-0* S0 seed stock, that likely results from the partial and sequential demethylation of the 5' UTR and promoter (figure 43), is particularly interesting regarding that last point. Although there could be sampling bias, those latter epivariants do not seem to be frequent in the worldwide set of accessions which could indicate that they are less stable than fully methylated/demethylated ones. Consistent with this hypothesis, some 'intermediate' epialleles segregating in the Col-0* seed stock showed slight reduction/increase of their methylation/expression level across the S1 and S2 generations (figure 43A-B, Publication Fig. 2). Because those variations could also be partially explained by biological/experimental variation it would be interesting to analyse Col-0* 'intermediate' epivariants over more generations, in several individual plants and by growing all the generations in the same experiment to reduce potential environmental variation.

Another very interesting question emerging from the Col-0* S0 seed stock is how all those variants appeared from an evolutionary point of view. They could be the result of one demethylation event (spontaneous or induced) in one plant (figure 44A) that could be assimilated to a kind of burst of epigenetic variation or from repeated methylation or demethylation events along genealogies (figure 44B-C). The first model (A) is strongly supported by the fact that we expect few generations of bulking between Col-0 and Col-0* S0 stock. The models B-C rely on many generations necessary to accumulate epigenetic variation, which is unlikely. But model B could somehow fit with the slight decrease of methylation observed in some Col-0* lines and model C could fit with the idea that lowly methylated epialleles get remethylated at

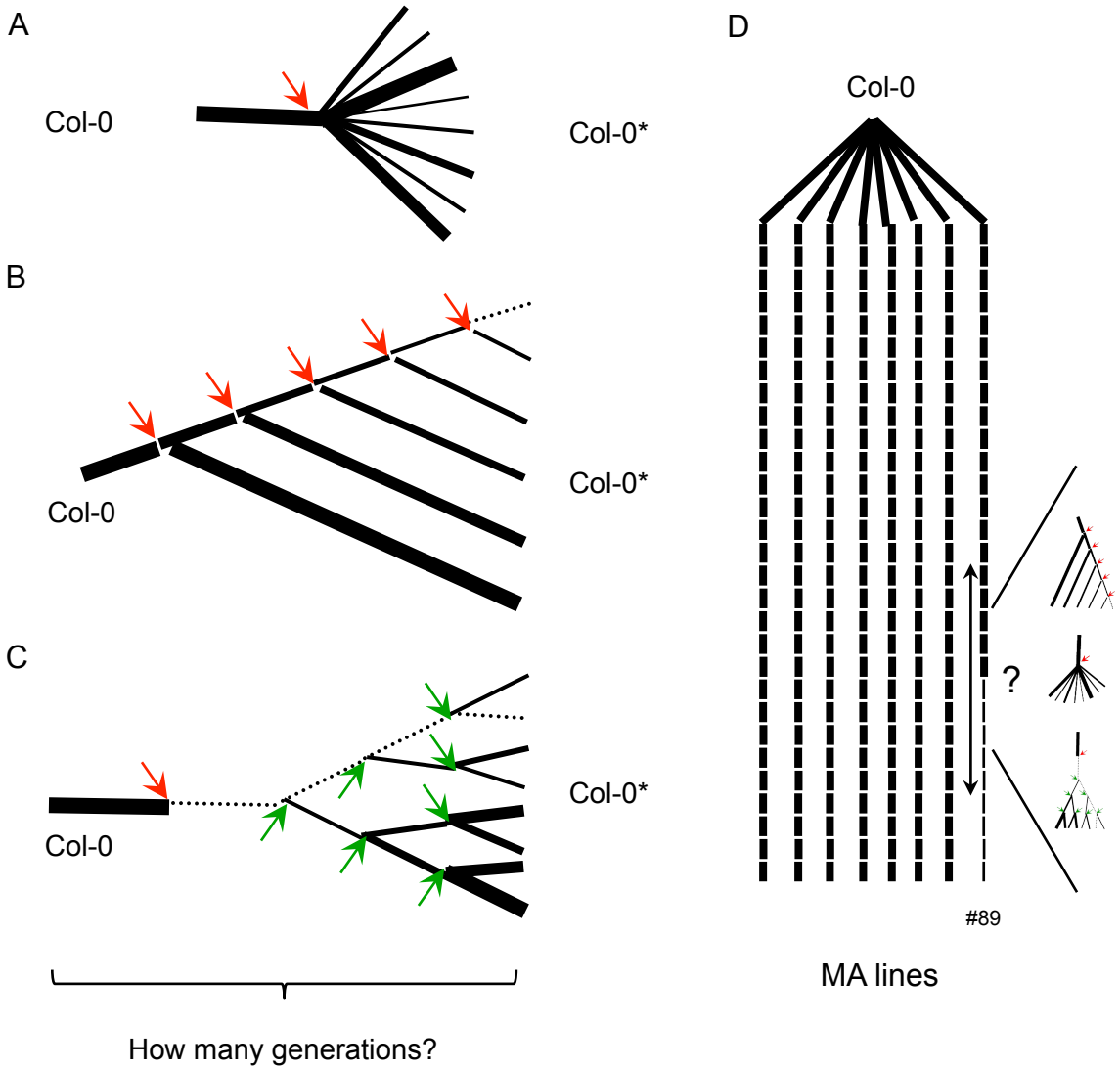


Fig. 44. Possible evolutionary pathways explaining the origin of the 'intermediate' *QQS* epivariants segregating in S0 Col-0* bulk. The thickness of each bar along genealogies is positively correlated to the level of *QQS* promoter methylation. 'Intermediate' *QQS* epivariants could arise during a single demethylation event in one plant (A) or from repeated demethylations along genealogies (B) or from a strong demethylation event followed by multiple remethylation (C). Analysing the parental lines of the MA#89 [64] could give information about how *QQS* epialleles appeared (D). Green and red arrows represent methylation and demethylation events respectively.

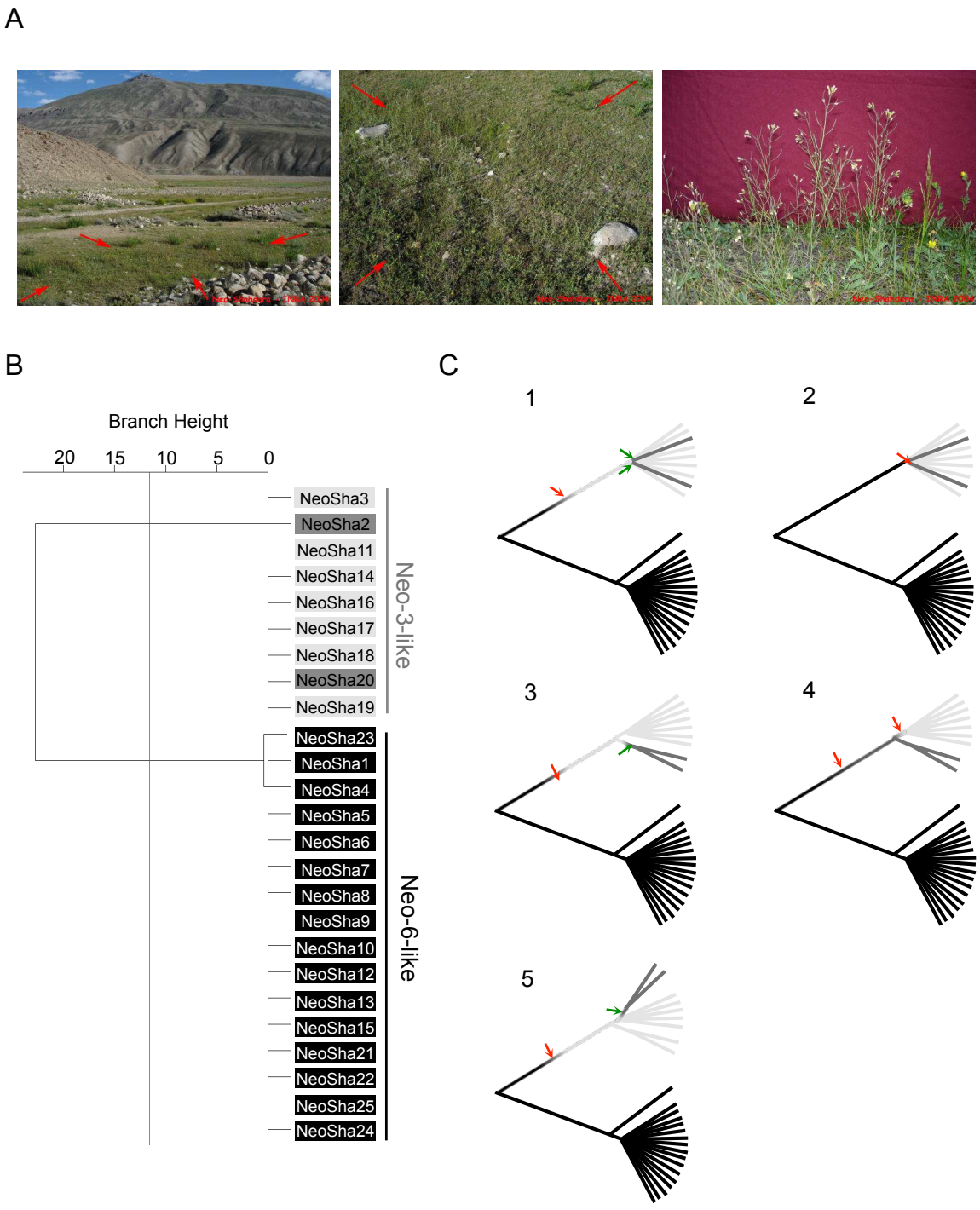


Fig. 45. *QQS* epialleles evolution in NeoSha population. A. Pictures of the NeoSha population collection site in Tajikistan at 3400m of altitude, in July 2004. Note that the size of the collection site was few square meters. B. A dendrogram obtained with AWclust from a matrix of 11 genotypic markers shows that two genetically distinct groups can be identified in Neo-Shahdara population. The red line cuts across the two main clusters identified. Plants highlighted in black or light grey are respectively methylated or demethylated in the 5' UTR of *QQS* gene. Likely 'intermediate' epialleles are highlighted in grey. C. Different possible evolutionary paths for *QQS* epialleles in NeoSha population. Green and red arrows represent methylation and demethylation events respectively. Because methylated alleles are more frequent that demethylated ones in *A. thaliana*, I supposed that the common ancestor of the NeoSha individuals has a methylated *QQS* epivariant.

some points (see above). Using the parental MA#89 lines (if available, [64]) for which we know the genealogy could be interesting to address the evolution of *QQS* epivariants by finding the generation(s) when *QQS* methylation switched and by looking at those generations to see if several 'intermediate' epivariants could be identified (figure 44D).

The NeoSha population sampled in Tajikistan (figure 45A) could also be particularly informative as 'intermediate' *QQS* epialleles are segregating in that set of individuals (Paper Fig4). Genetic clustering of the whole NeoSha population available (25 individuals) using 11 microsatellites reveals two genetic subgroups. All members of the 'Neo-6-like'cluster (i.e. 16 individuals) have a methylated 5' UTR *QQS* epiallele whereas all members of the 'Neo-3-like'cluster (9 individuals) except two (NeoSha-2 – 'intermediate' epiallele – and NeoSha-20), have an unmethylated 5' UTR epiallele (figure 45B). Note that all individuals of the NeoSha population are completely methylated in the promoter of *QQS* (figure 43B). Several evolutive scenarios can be imagined to explain the level of *QQS* methylation in that population but differentiating between them requires more information about the genealogies of Neo-3-like and to a lesser extent Neo-6-like subgroups. The sequencing of some individuals in the two groups in collaboration with Detlef Weigel will be a first step to see if reconstructing such genealogies using sequence data is possible (depending on the genetic distance between these individuals and the time-scale of their evolution). Another way could be to analyse the methylome of those individuals. Indeed as mentioned in the introduction, genome-wide methylation variants are more frequent than genetic mutations and could potentially perform better than genetic variants to reconstruct recent genealogies. If no particular genealogies in each subgroup can be distinguished, this could suggest that all Neo-3-like (or Neo-6-like) accessions are issued from a single individual. This would be consistent with the hypothesis that bursts of epigenetic diversity can occur at the level of a single individual (figure 45C – models 1 and 2). If a genealogy of the Neo-3-like group can be obtained and the two demethylated alleles grouped together at the base of the genealogy, it could suggest that an independent event is at the source of methylation contrast between these two individuals and the rest of the Neo-3-like (figure 45C – models 3 and 4). Whereas the former possible observations would not allow to infer the ancestral state (methylated or demethylated) of the Neo-3-like common closer ancestor, if the two demethylated alleles grouped together within the genealogy, it would strongly suggest that the common closer ancestor of Neo-3-like accessions was demethylated and that NeoSha-2 and NeoSha-20 'intermediate' epialleles have been remethylated (figure 45C – models 5) providing a clear view of the evolution of *QQS* epivariants in that natural population.

In the previous model, I made the assumption that the ancestral allele of Neo-3-like and Neo-6-like subgroups was methylated because methylated alleles are more frequent at the world-wide scale, but this assumption still need to be proven. Because NeoSha individuals presumably represent the direct descendants of Shahdara, this accession, likely methylated at the 5' UTR

Tab. 7. Genetic diversity in Shahdara and NeoSha individuals.

Reference marker name	Chromosome	TAIR10	SNP	Col-0	Sha	Neo-6	Neo-23	Neo-3
mk10008	1	2211034	A/G	G	G	G	A	G
T27K12	1	15926702	msat		NA	b	b	h
mk10072	1	24187552	A/G	G	A	G	A	A
mk20102	2	764529	T/C	T	T	T	T	C
mk20103	2	764693	T/C	A	A	A	A	G
mk20105	2	1427380	T/A	A	NA	T	T	A
MSAT2.26	2	1945453	msat		h	h	h	b
mk20119	2	7544501	T/G	C	A	A	C	C
MSAT2.4	2	13831870	msat		NA	b	b	h
NGA172	3	786303	msat		h	h	h	b
mk30167	3	2231537	T/G	G	T	G	G	T
mk30168	3	2231932	T/C	A	G	A	A	G
Msat3.19	3	8808167	msat		h	b	NA	h
Msat3.1	3	12170600	msat		h	h	h	b
QQS	**3**	**12349047**			**me**	**me**	**me**	**deme**
Msat3.21	3	17282665	msat		b	h	h	NA
mk40235	4	1055234	T/C	A	A	G	G	A
mk40236	4	1055329	T/C	T	T	C	C	T
NGA8	4	5628810	msat		b	h	h	b
mk40255	4	8575595	A/C	C	A	C	A	A
Mast4.18	4	11966304	msat		b	h	h	b
mk40281	4	17038397	A/T	A	T	T	T	A
mk50318	5	6523118	A/G	T	T	C	C	T
ICE5	5	7705251	msat		NA	h	b	b
mk50334	5	9723212	G/C	C	C	G	C	C
mk50344	5	15047966	A/G	A	A	G	A	A
mk50346	5	15750717	T/C	A	NA	G	A	A
ICE3	5	16144941	msat		h	h	h	b
mk50349	5	16351105	T/C	T	T	T	T	C
mk50350	5	16428797	C/G	C	C	C	C	G
JV75/76	5	23879358	msat		NA	b	NA	h
MUM2	5	25530055	msat		120	162	162	120
Cytotype	chloroplast and mitochondria markers							

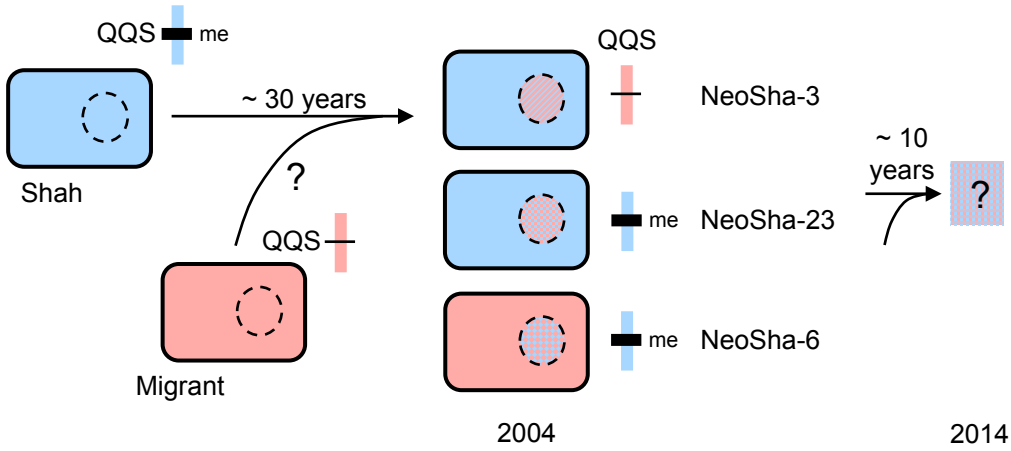


Fig. 46. Possible origin of NeoSha population. NeoSha population could arise from an event of hybridization of Shah descendants with a migrant accession resulting in individuals with the Shah or migrant cytoplasms and a mosaic of both nuclear genome. Because we do not detect heterozygous loci, the hybridization process is likely to be relatively old allowing the fixation of the nuclear genomes. Such evolutive history could explain the segregation of the *QQS* epivariants, the methylated allele being inherited from Shahdara and the demethylated one from the migrant population. It would be interesting to go back in Tadjikistan in 2014 to analyse how the NeoSha population evolved.

of *QQS* gene (Publication – Fig. 4B), could be used to define the ancestral state of *QQS* in NeoSha population. Nevertheless, the recent genotyping of Shah, NeoSha-3, NeoSha-6 and NeoSha-23 (including cytotype, table 7) showed that the evolution of this population might be complicated and may involve hybridization within the ancestral local population (if Shah) and/or with a migrant (figure 46), especially to explain the presence of a distinct cytoplasm. Considering this scenario, the methylated and demethylated epialleles segregating in NeoSha population could be relatively ancient. In 2014, it will be 10 years since Olivier collected the NeoSha population we are currently working on. Going back there to perform an 'extensive' sampling of the 'NeoSha' population -and possibly other populations around this area- could provide precious information about the evolution of *QQS* epialleles in terms of stability and frequency. Alternatively, the different materials available from this project could be used as starting material to experimentally assess the stability of the different epialleles under various (controlled) environmental conditions, especially stress-inducing environments.

CONCLUSION

During my PhD work I was interested in understanding how genetic ('MOT1' and 'EGM' projects) and epigenetic ('QQS' project) diversity shape phenotypic variation in natural accessions of *A. thaliana* under various environmental conditions. More particularly, I contributed to the identification and characterization of a new allele of a Mo transporter –*MOT1[Sha]*– potentially important regarding *A. thaliana* adaptation to soils rich in Mo. This study highlights the importance of genetic heterogeneity in shaping *A. thaliana* natural variation, a characteristic of trait's architecture to which we should bring more attention, particularly regarding GWA analyses that are more and more used in *A. thaliana* as well as in non-model and/or cultivated species. This project also stresses the precious informations that could bring a better characterization of the soils and more generally the ecological and environmental conditions where accessions are growing. Then, in the 'EGM' project I contributed to the characterization of a new pathway that could link part of the phenotype observed under *in vitro* mannitol condition to a defense mechanism against mannitol producing pathogens. This work highlight the potential of natural variation analyses for the identification of new genes involved in stress responses although in that case, the identified variants are only indirectly related to the initially-intended screen. It also underlines the importance of RLKs family expansion for plant response to environmental cues. From my point of view, the two closely identified *EGM1* and *EGM2* paralogs could be a good starting point for the understanding of the molecular basis of RLK subfunctionalisation. Finally, I participated to the characterization of an epialleles segregating at a relatively high frequency in natural population of *A. thaliana*. Although the ecological importance of QQS epialleles is not clear, our work highlights the contribution of epigenetic variations to change in transcript accumulation in natural populations. It remains to be determined in more details how much heritable variations in DNA methylation account for the variations in transcript accumulation observed in natural accessions [180, 181, 41, 42].

Regarding, the interaction between natural alleles, phenotypic traits and environmental variables, all the projects I worked on highlight several difficulties that can be faced when trying to associate genetic mutation to phenotypic variation. First, identifying a phenotype is not always straightforward. For example, we could not find a phenotype associated with *QQS* natural epivariants even though it is known from the bibliography that 'artificial' *QQS* mutants and overexpressors are perturbed for starch metabolism [341, 344]. Because genetic

and epigenetic variants could have an effect in a particular level of biological organization, in a given organ, at a given developmental stage or under a specific environment, special care must be taken when stipulating that they have no phenotypic effect. Conversely, it is not because a variant has a phenotypic effect in one condition (internal and/or external) that this phenotype holds in others. Particularly, most of transcriptomic and metabolic variations are likely to be buffered at the developmental and morphological level [177]. Although huge progresses have been made in the understanding of phenotypic networks, estimating the impact of genetic and epigenetic mutations on the global plant phenotype, even in a given environment, will probably remain challenging for the next decades.

Then, the 'EGM' project highlights that some confusion could arise regarding the nature of the environmental variables responsible for a phenotype. Indeed, mannitol was for long used *in vitro* to induce an osmotic stress but our data suggest that the molecule by itself could result in a specific response that may be associated to pathogen defence. It does not mean that all the studies that used mannitol as an osmoticum were wrong because at high concentration (300mM) the major effect of mannitol is probably an osmotic constraint. But it points out that the pathways associated with mannitol response likely depend on the concentration of mannitol used *in vitro* and we can't exclude that synergistic effects between the two pathways arise at intermediate concentrations. Similarly, distinguishing the effect of osmotic stress from the one of Na+ toxicity can be problematic for people working on salt stress [348]. We show that working on *in vitro* media ('EGM') or artificial soils ('MOT1') can help discovering new pathways, new genes and/or new alleles associated to a trait. However, when interested in adaptation, we can wonder if those phenotyping methods are appropriate. Indeed, flowering QTLs isolated in greenhouse conditions and in semi-natural settings are quite different [141] and differences are likely to be more important with QTLs isolated *in vitro*. Overall, because much of the QTLs isolated so far correspond to 'artificial' conditions QTLs, we can wonder how much our knowledge about *A. thaliana* adaptation is biased toward the identification of alleles constrained by environment variables easy to control artificially (light, temperature). Are those climatic factors really major for *A. thaliana* adaptation compared to biotic and abiotic parameters? Besides, what is the role of intraspecific competition in the distribution of *A. thaliana*?

Finally, in the 'MOT1' and 'EGM' projects, we show that the benefit of new alleles can vary depending on environmental parameters (here Mo availability and pathogens' presence) so that selection processes (positive or negative) are likely to act in specific but not all environments. Thus the results obtained during this PhD comfort the idea that conditional neutrality is an important factor for the occurrence and maintenance of genetic diversity in *A. thaliana* populations [211]. At another scale, we can wonder if conditional neutrality also drives the evolution of new-born genes. This is a reasonable hypothesis as those later genes, when they

appear, are likely not essential–at least in most environments–. But they may become particularly important under specific constraints. Our work on *QQS* epialleles suggests that epigenetic variation may contribute to the evolution of new-born genes by enabling them to adjust their expression in a rapid, heritable but reversible manner. However it remains to be determined if those adjustments are dependent on environmental constraints or not and if they have an ecological significance.

APPENDIX

A. ADDITIONAL PUBLICATION

A review was written and published in Current Opinion in Plant Biology in 2012 but cannot be displayed in this online version of my manuscript for copyright reasons.

BIBLIOGRAPHY

[1] Al-Shehbaz, I.A. and O'Kane Jr, S.L., *Taxonomy and phylogeny of Arabidopsis (brassicaceae)*, The Arabidopsis Book **6**, 1-22 (2002). 19, 21

[2] Clauss, M. and Koch, M., *Poorly known relatives of Arabidopsis thaliana.*, Trends Plant Sci **11(9)**, 449-459 (2006). 19

[3] Hunter, B. and Bomblies, K., *Progress and promise in using Arabidopsis to study adaptation, divergence, and speciation.*, Arabidopsis Book **8**, e0138 (2010). 19, 21

[4] Platt, A., Horton, M., Huang, Y., Li, Y., Anastasio, A., Mulyati, N., Agren, J., Bossdorf, O., Byers, D., Donohue, K., Dunning, M., Holub, E., Hudson, A., Le Corre, V., Loudet, O., Roux, F., Warthmann, N., Weigel, D., Rivero, L., Scholl, R., Nordborg, M., Bergelson, J. and Borevitz, J., *The scale of population structure in Arabidopsis thaliana.*, PLoS Genet **6(2)**, e1000843 (2010). 20, 24, 25, 26

[5] Bomblies, K., Yant, L., Laitinen, R., Kim, S., Hollister, J., Warthmann, N., Fitz, J. and Weigel, D., *Local-scale patterns of genetic variability, outcrossing, and spatial structure in natural stands of Arabidopsis thaliana.*, PLoS Genet **6(3)**, e1000890 (2010). 20, 25, 26

[6] Charlesworth, D. and Vekemans, X., *How and when did Arabidopsis thaliana become highly self-fertilising.*, Bioessays **27(5)**, 472-476 (2005). 20

[7] Hu, T., Pattyn, P., Bakker, E., Cao, J., Cheng, J., Clark, R., Fahlgren, N., Fawcett, J., Grimwood, J., Gundlach, H., Haberer, G., Hollister, J., Ossowski, S., Ottilar, R., Salamov, A., Schneeberger, K., Spannagl, M., Wang, X., Yang, L., Nasrallah, M., Bergelson, J., Carrington, J., Gaut, B., Schmutz, J., Mayer, K., Van De Peer, Y., Grigoriev, I., Nordborg, M., Weigel, D. and Guo, Y., *The Arabidopsis lyrata genome sequence and the basis of rapid genome size change.*, Nature Genetics **43(5)**, 476-481 (2011). 20, 21

[8] Hoffmann, M., *Biogeography of Arabidopsis thaliana (l.) heynh.(Brassicaceae)*, Journal of Biogeography **29(1)**, 125-134 (2002). 20, 79

[9] Hoffmann, M. and Soltis, P., *Evolution of the realized climatic niche in the genus Arabidopsis (Brassicaceae)*, Evolution **59(7)**, 1425-1436 (2005). 21

[10] Shindo, C., Bernasconi, G. and Hardtke, C., *Natural genetic variation in Arabidopsis: Tools, traits and prospects for evolutionary ecology.*, Ann Bot **99(6)**, 1043-1054 (2007). 21, 22

[11] Montesinos, A., Tonsor, S., Alonso-Blanco, C. and Pico, F., *Demographic and genetic patterns of variation among populations of Arabidopsis thaliana from contrasting native environments.*, PLoS One **4(9)**, e7213 (2009). 23, 25, 26, 68

[12] Weigel, D. and Mott, R., *The 1001 genomes project for Arabidopsis thaliana.*, Genome Biol **10(5)**, 107 (2009). 24

[13] Budar, F. and Roux, F., *The role of organelle genomes in plant adaptation: Time to get to work!*, Plant Signal Behav **6(5)**, 635-639 (2011). 24

[14] Nordborg, M., Hu, T., 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, N., Shah, C., Wall, J., Wang, J., Zhao, K., Kalbfleisch, T., Schulz, V., Kreitman, M. and Bergelson, J., *The pattern of polymorphism in Arabidopsis thaliana.*, PLoS Biol **3(7)**, e196 (2005). 25, 26, 29, 50, 51, 75, 113

[15] Schmid, K., Törjék, O., Meyer, R., Schmuths, H., Hoffmann, M. and Altmann, T., *Evidence for a large-scale population structure of Arabidopsis thaliana from genome-wide single nucleotide polymorphism markers*, Theoretical and Applied Genetics **112(6)**, 1104-1114 (2006). 25, 26, 29

[16] Francois, O., Blum, M., Jakobsson, M. and Rosenberg, N., *Demographic history of european populations of Arabidopsis thaliana.*, PLoS Genet **4(5)**, e1000075 (2008). 26, 113

[17] Sharbel, T., Haubold, B. and Mitchell-Olds, T., *Genetic isolation by distance in Arabidopsis thaliana: Biogeography and postglacial colonization of Europe*, Mol Ecol **9(12)**, 2109-2118 (2000). 25, 26

[18] Horton, M., Hancock, A., Huang, Y., Toomajian, C., Atwell, S., Auton, A., Muliyati, N., Platt, A., Sperone, F., Vilhjalmsson, B., Nordborg, M., Borevitz, J. and Bergelson, J., *Genome-wide patterns of genetic variation in worldwide Arabidopsis thaliana accessions from the RegMap panel.*, Nat Genet **44(2)**, 212-216 (2012). 25, 26, 50, 76, 78, 79

[19] Cao, J., Schneeberger, K., Ossowski, S., Gunther, T., Bender, S., Fitz, J., Koenig, D., Lanz, C., Stegle, O., Lippert, C., Wang, X., Ott, F., Muller, J., Alonso-Blanco, C., Borgwardt, K., Schmid, K. and Weigel, D., *Whole-genome sequencing of multiple Arabidopsis thaliana populations*, Nat Genet **43(10)**, 956-963 (2011). 25, 26, 28, 29, 51, 75, 77, 113

[20] Abbott, R. and Gomes, M., *Population genetic structure and outcrossing rate of Arabidopsis thaliana (l.) heynh*, Heredity **62(Pt 3)**, 411-418 (1989). 25

[21] Hanfstingl, U., Berry, A., Kellogg, E., Costa, J.R., Rudiger, W. and Ausubel, F., *Haplotypic divergence coupled with lack of diversity at the Arabidopsis thaliana alcohol dehydrogenase locus: Roles for both balancing and directional selection?*, Genetics **138(3)**, 811-828 (1994). 25

[22] Innan, H., Tajima, F., Terauchi, R. and Miyashita, N., *Intragenic recombination in the Adh locus of the wild plant Arabidopsis thaliana.*, Genetics **143(4)**, 1761-1770 (1996). 25

[23] Kawabe, A., Innan, H., Terauchi, R. and Miyashita, N., *Nucleotide polymorphism in the acidic chitinase locus (Chia) region of the wild plant Arabidopsis thaliana.*, Mol Biol Evol **14(12)**, 1303-1315 (1997). 25

[24] Innan, H., Terauchi, R. and Miyashita, N., *Microsatellite polymorphism in natural populations of the wild plant Arabidopsis thaliana.*, Genetics **146(4)**, 1441-1452 (1997). 25

[25] Purugganan, M. and Suddith, J., *Molecular population genetics of the Arabidopsis CAULIFLOWER regulatory gene: Nonneutral evolution and naturally occurring variation in floral homeotic function.*, Proc Natl Acad Sci U S A **95(14)**, 8130-8134 (1998). 25

[26] Bergelson, J., Stahl, E., Dudek, S. and Kreitman, M., *Genetic variation within and among populations of Arabidopsis thaliana.*, Genetics **148(3)**, 1311-1323 (1998). 25

[27] Miyashita, N., Kawabe, A. and Innan, H., *DNA variation in the wild plant Arabidopsis thaliana revealed by amplified fragment length polymorphism analysis.*, Genetics **152(4)**, 1723-1731 (1999). 25

[28] Zwan, C.V., Brodie, S.A. and Campanella, J.J., *The intraspecific phylogenetics of Arabidopsis thaliana in worldwide populations*, Systematic Botany **25**, 47-59 (2000). 25

[29] Barth, S., Melchinger, A. and Lubberstedt, T., *Genetic diversity in Arabidopsis thaliana l. Heynh. Investigated by cleaved amplified polymorphic sequence (CAPS) and inter-simple sequence repeat (ISSR) markers.*, Mol Ecol **11(3)**, 495-505 (2002). 25

[30] Schmid, K., Sorensen, T., Stracke, R., Torjek, O., Altmann, T., Mitchell-Olds, T. and Weisshaar, B., *Large-scale identification and analysis of genome-wide single-nucleotide polymorphisms for mapping in Arabidopsis thaliana.*, Genome Res **13(6A)**, 1250-1257 (2003). 25

[31] Jorgensen, S. and Mauricio, R., *Neutral genetic variation among wild North American populations of the weedy plant Arabidopsis thaliana is not geographically structured.*, Mol Ecol **13(11)**, 3403-3413 (2004). 25

[32] Mckhann, H., Camilleri, C., Berard, A., Bataillon, T., David, J., Reboud, X., Le Corre, V., Caloustian, C., Gut, I. and Brunel, D., *Nested core collections maximizing genetic diversity in Arabidopsis thaliana.*, Plant J **38(1)**, 193-202 (2004). 25

[33] Stenoien, H., Fenster, C., Tonteri, A. and Savolainen, O., *Genetic variability in natural populations of Arabidopsis thaliana in northern Europe.*, Mol Ecol **14(1)**, 137-148 (2005). 25

[34] Ostrowski, M., David, J., Santoni, S., Mckhann, H., Reboud, X., Le Corre, V., Camilleri, C., Brunel, D., Bouchez, D., Faure, B. and Bataillon, T., *Evidence for a large-scale population structure among accessions of Arabidopsis thaliana: Possible causes and consequences for the distribution of linkage disequilibrium.*, Mol Ecol **15(6)**, 1507-1517 (2006). 25

[35] Clark, R., Schweikert, G., Toomajian, C., Ossowski, S., Zeller, G., Shinn, P., Warthmann, N., Hu, T., Fu, G., Hinds, D., Chen, H., Frazer, K., Huson, D., Scholkopf, B., Nordborg, M., Ratsch, G., Ecker, J. and Weigel, D., *Common sequence polymorphisms shaping genetic diversity in Arabidopsis thaliana.*, Science **317(5836)**, 338-342 (2007). 25, 28, 29, 75

[36] He, F., Kang, D., Ren, Y., Qu, L., Zhen, Y. and Gu, H., *Genetic diversity of the natural populations of Arabidopsis thaliana in China.*, Heredity (Edinb) **99(4)**, 423-431 (2007). 25

[37] Beck, J., Schmuths, H. and Schaal, B., *Native range genetic variation in Arabidopsis thaliana is strongly geographically structured and reflects Pleistocene glacial dynamics.*, Mol Ecol **17(3)**, 902-915 (2008). 25

[38] Pico, F., Mendez-Vigo, B., Martinez-Zapater, J.M. and Alonso-Blanco, C., *Natural genetic variation of Arabidopsis thaliana is geographically structured in the Iberian peninsula.*, Genetics **180(2)**, 1009-1021 (2008). 25

[39] Yin, P., Kang, J., He, F., Qu, L. and Gu, H., *The origin of populations of Arabidopsis thaliana in China, based on the chloroplast DNA sequences.*, BMC Plant Biol **10(1)**, 22 (2010). 25

[40] Moison, M., Roux, F., Quadrado, M., Duval, R., Ekovich, M., Le, D., Verzaux, M. and Budar, F., *Cytoplasmic phylogeny and evidence of cyto-nuclear co-adaptation in Arabidopsis thaliana.*, Plant J **63(5)**, 728-738 (2010). 25

[41] Gan, X., Stegle, O., Behr, J., Steffen, J., Drewe, P., Hildebrand, K., Lyngsoe, R., Schultheiss, S., Osborne, E., Sreedharan, V., Kahles, A., Bohnert, R., Jean, G., Derwent, P., Kersey, P., Belfield, E., Harberd, N., Kemen, E., Toomajian, C., Kover, P., Clark, R., Ratsch, G. and Mott, R., *Multiple reference genomes and transcriptomes for Arabidopsis thaliana.*, Nature **477(7365)**, 419-423 (2011). 25, 28, 29, 49, 196, 203

[42] Schmitz, R., Schultz, M., Urich, M., Nery, J., Pelizzola, M., Libiger, O., Alix, A., Mccosh, R., Chen, H., Schork, N. and Ecker, J., *Patterns of population epigenomic diversity.*, Nature **495(7440)**, 193-198 (2013). 25, 32, 33, 49, 203

[43] Ossowski, S., Schneeberger, K., Lucas-Lledo, J.I., Warthmann, N., Clark, R., Shaw, R., Weigel, D. and Lynch, M., *The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana.*, Science **327(5961)**, 92-94 (2010). 27, 28, 33, 196

[44] Marriage, T., Hudman, S., Mort, M., Orive, M., Shaw, R. and Kelly, J., *Direct estimation of the mutation rate at dinucleotide microsatellite loci in Arabidopsis thaliana (brassicaceae).*, Heredity (Edinb) **103(4)**, 310-317 (2009). 27

[45] Schneeberger, K., Ossowski, S., Ott, F., Klein, J., Wang, X., Lanz, C., Smith, L., Cao, J., Fitz, J., Warthmann, N., Henz, S., Huson, D. and Weigel, D., *Reference-guided assembly of four diverse Arabidopsis thaliana genomes.*, Proc Natl Acad Sci U S A **108(25)**, 10249-10254 (2011). 28, 29

[46] Zhang, X., Shiu, S., Cal, A. and Borevitz, J., *Global analysis of genetic, epigenetic and transcriptional polymorphisms in Arabidopsis thaliana using whole genome tiling arrays.*, PLoS Genet **4(3)**, e1000032 (2008). 29, 32

[47] Schmuths, H., Meister, A., Horres, R. and Bachmann, K., *Genome size variation among accessions of Arabidopsis thaliana.*, Ann Bot **93(3)**, 317-321 (2004). 29

[48] Davison, J., Tyagi, A. and Comai, L., *Large-scale polymorphism of heterochromatic repeats in the DNA of Arabidopsis thaliana.*, BMC Plant Biol **7**, 44 (2007). 29

[49] Ziolkowski, P., Koczyk, G., Galganski, L. and Sadowski, J., *Genome sequence comparison of Col and Ler lines reveals the dynamic nature of Arabidopsis chromosomes.*, Nucleic Acids Res **37(10)**, 3189-3201 (2009). 29

[50] Boyko, A. and Kovalchuk, I., *Genome instability and epigenetic modification–heritable responses to environmental stress?*, Curr Opin Plant Biol **14(3)**, 260-266 (2011). 29, 30

[51] Mirouze, M. and Paszkowski, J., *Epigenetic contribution to stress adaptation in plants.*, Curr Opin Plant Biol **14(3)**, 267-274 (2011). 29, 30

[52] Boyko, A., Kathiria, P., Zemp, F., Yao, Y., Pogribny, I. and Kovalchuk, I., *Transgenerational changes in the genome stability and methylation in pathogen-infected plants: (virus-induced plant genome instability).*, Nucleic Acids Res **35(5)**, 1714-1725 (2007). 29

[53] Debolt, S., *Copy number variation shapes genome diversity in Arabidopsis over immediate family generational scales.*, Genome Biol Evol **2**, 441-453 (2010). 29

[54] Lister, R., O'Malley, R.C., Tonti-Filippini, J., Gregory, B., Berry, C., Millar, A. and Ecker, J., *Highly integrated single-base resolution maps of the epigenome in Arabidopsis.*, Cell **133(3)**, 523-536 (2008). 30

[55] Cokus, S., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C., Pradhan, S., Nelson, S., Pellegrini, M. and Jacobsen, S., *Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning.*, Nature **452(7184)**, 215-219 (2008). 30

[56] Zhang, X., Yazaki, J., Sundaresan, A., Cokus, S., Chan, S., Chen, H., Henderson, I., Shinn, P., Pellegrini, M., Jacobsen, S. and Ecker, J., *Genome-wide high-resolution mapping and functional analysis of DNA methylation in Arabidopsis.*, Cell **126(6)**, 1189-1201 (2006). 30

[57] Zilberman, D., Gehring, M., Tran, R., Ballinger, T. and Henikoff, S., *Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription.*, Nat Genet **39(1)**, 61-69 (2007). 30

[58] Richards, E., *Inherited epigenetic variation–revisiting soft inheritance.*, Nat Rev Genet **7(5)**, 395-401 (2006). 30

[59] Woo, H., Pontes, O., Pikaard, C. and Richards, E., *VIM1, a methylcytosine-binding protein required for centromeric heterochromatinization.*, Genes Dev **21(3)**, 267-277 (2007). 30, 31

[60] Liu, J., He, Y., Amasino, R. and Chen, X., *SiRNAs targeting an intronic transposon in the regulation of natural flowering behavior in Arabidopsis.*, Genes Dev **18(23)**, 2873-2878 (2004). 30, 56

[61] Fujimoto, R., Sasaki, T., Kudoh, H., Taylor, J., Kakutani, T. and Dennis, E., *Epigenetic variation in the FWA gene within the genus Arabidopsis.*, Plant J **66(5)**, 831-843 (2011). 30

[62] Durand, S., Bouche, N., Perez Strand, E., Loudet, O. and Camilleri, C., *Rapid establishment of genetic incompatibility through natural epigenetic variation.*, Curr Biol **22(4)**, 326-331 (2012). 30, 31, 32, 54, 59

[63] Vaughn, M., Tanurdzic, M., Lippman, Z., Jiang, H., Carrasquillo, R., Rabinowicz, P., Dedhia, N., Mccombie, W., Agier, N., Bulski, A., Colot, V., Doerge, R. and Martienssen, R., *Epigenetic natural variation in Arabidopsis thaliana.*, PLoS Biol **5(7)**, e174 (2007). 32

[64] Becker, C., Hagmann, J., Muller, J., Koenig, D., Stegle, O., Borgwardt, K. and Weigel, D., *Spontaneous epigenetic variation in the Arabidopsis thaliana methylome.*, Nature **480(7376)**, 245-249 (2011). 32, 33, 197, 199

[65] Schmitz, R., Schultz, M., Lewsey, M., O'Malley, R.C., Urich, M., Libiger, O., Schork, N. and Ecker, J., *Transgenerational epigenetic instability is a source of novel methylation variants.*, Science **334(6054)**, 369-373 (2011). 32, 33

[66] Mirouze, M., Reinders, J., Bucher, E., Nishimura, T., Schneeberger, K., Ossowski, S., Cao, J., Weigel, D., Paszkowski, J. and Mathieu, O., *Selective epigenetic control of retrotransposition in Arabidopsis.*, Nature **461(7262)**, 427-430 (2009). 32

[67] Becker, C. and Weigel, D., *Epigenetic variation: Origin and transgenerational inheritance.*, Curr Opin Plant Biol **15(5)**, 562-567 (2012). 32, 33

[68] Dong, X., Reimer, J., Gobel, U., Engelhorn, J., He, F., Schoof, H. and Turck, F., *Natural variation of H3K27me3 distribution between two Arabidopsis accessions and its association with flanking transposable elements.*, Genome Biol **13(12)**, R117 (2012). 33

[69] Pecinka, A., Dinh, H., Baubec, T., Rosa, M., Lettner, N. and Mittelsten Scheid, O., *Epigenetic regulation of repetitive elements is attenuated by prolonged heat stress in Arabidopsis.*, Plant Cell **22(9)**, 3118-3129 (2010). 34

[70] Tittel-Elmer, M., Bucher, E., Broger, L., Mathieu, O., Paszkowski, J. and Vaillant, I., *Stress-induced activation of heterochromatic transcription.*, PLoS Genet **6(10)**, e1001175 (2010). 34

[71] Molinier, J., Ries, G., Zipfel, C. and Hohn, B., *Transgeneration memory of stress in plants.*, Nature **442(7106)**, 1046-1049 (2006). 34

[72] Paszkowski, J. and Grossniklaus, U., *Selected aspects of transgenerational epigenetic inheritance and resetting in plants.*, Curr Opin Plant Biol **14(2)**, 195-203 (2011). 34

[73] Wilczek, A., Burghardt, L., Cobb, A., Cooper, M., Welch, S. and Schmitt, J., *Genetic and physiological bases for phenological responses to current and predicted climates.*, Philos Trans R Soc Lond B Biol Sci **365(1555)**, 3129-3147 (2010). 35, 66, 78, 79

[74] Atwell, S., Huang, Y., Vilhjalmsson, B., Willems, G., Horton, M., Li, Y., Meng, D., Platt, A., Tarone, A., Hu, T., Jiang, R., Muliyati, N., Zhang, X., Amer, M., Baxter, I., Brachi, B., Chory, J., Dean, C., Debieu, M., De Meaux, J., Ecker, J., Faure, N., Kniskern, J., Jones, J., Michael, T., Nemri, A., Roux, F., Salt, D., Tang, C., Todesco, M., Traw, M., Weigel, D., Marjoram, P., Borevitz, J., Bergelson, J. and Nordborg, M., *Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines.*, Nature **465(7298)**, 627-631 (2010). 35, 50, 51, 56, 57, 86

[75] Granier, C., Aguirrezabal, L., Chenu, K., Cookson, S., Dauzat, M., Hamard, P., Thioux, J., Rolland, G., Bouchier-Combaud, S., Lebaudy, A., Muller, B., Simonneau, T. and Tardieu, F., *PHENOPSIS, an automated platform for reproducible phenotyping of plant responses to soil water deficit in Arabidopsis thaliana permitted the identification of an accession with low sensitivity to soil water deficit.*, New Phytol **169(3)**, 623-635 (2006). 35

[76] Tisne, S., Serrand, Y., Bach, L., Gilbault, E., Ben Ameur, R., Balasse, H., Voisin, R., Bouchez, D., Durand-Tardif, M., Guerche, P., Chareyron, G., Da Rugna, J., Camilleri, C. and Loudet, O., *PHENOSCOPE: An automated large-scale phenotyping platform offering high spatial homogeneity.*, Plant J **74(3)**, 534-544 (2013). 35, 192

[77] Joosen, R., Ligterink, W., Hilhorst, H. and Keurentjes, J., *Advances in genetical genomics of plants.*, Curr Genomics **10(8)**, 540-549 (2009). 35

[78] Bylesjo, M., Segura, V., Soolanayakanahally, R., Rae, A., Trygg, J., Gustafsson, P., Jansson, S. and Street, N., *LAMINA: A tool for rapid quantification of leaf size and shape parameters.*, BMC Plant Biol **8**, 82 (2008). 35

[79] Backhaus, A., Kuwabara, A., Bauch, M., Monk, N., Sanguinetti, G. and Fleming, A., *LEAF-PROCESSOR: A new leaf phenotyping tool using contour bending energy and shape cluster analysis.*, New Phytol **187(1)**, 251-261 (2010). 35

[80] Armengaud, P., *Ez-Rhizo software: The gateway to root architecture analysis.*, Plant Signal Behav **4(2)**, 139-141 (2009). 35

[81] French, A., Ubeda-Tomas, S., Holman, T., Bennett, M. and Pridmore, T., *High-throughput quantification of root growth using a novel image-analysis tool.*, Plant Physiol **150(4)**, 1784-1795 (2009). 35

[82] Jansen, M., Martret, B. and Koornneef, M., *Variations in constitutive and inducible UV-B tolerance; dissecting photosystem ii protection in Arabidopsis thaliana accessions.*, Physiol Plant **138(1)**, 22-34 (2009). 35

[83] De Vos, R.C., Moco, S., Lommen, A., Keurentjes, J., Bino, R. and Hall, R., *Untargeted large-scale plant metabolomics using liquid chromatography coupled to mass spectrometry.*, Nat Protoc **2(4)**, 778-791 (2007). 37

[84] Lisec, J., Schauer, N., Kopka, J., Willmitzer, L. and Fernie, A., *Gas chromatography mass spectrometry-based metabolite profiling in plants.*, Nat Protoc **1(1)**, 387-396 (2006). 37

[85] Keurentjes, J., *Genetical metabolomics: Closing in on phenotypes.*, Curr Opin Plant Biol **12(2)**, 223-230 (2009). 37

[86] Lahner, B., Gong, J., Mahmoudian, M., Smith, E., Abid, K., Rogers, E., Guerinot, M., Harper, J., Ward, J., Mcintyre, L., Schroeder, J. and Salt, D., *Genomic scale profiling of nutrient and trace elements in Arabidopsis thaliana.*, Nat Biotechnol **21(10)**, 1215-1221 (2003). 37

[87] Salt, D., Baxter, I. and Lahner, B., *Ionomics and the study of the plant ionome.*, Annu Rev Plant Biol **59**, 709-733 (2008). 37

[88] Weigel, D., *Natural variation in Arabidopsis: From molecular genetics to ecological genomics.*, Plant Physiol **158(1)**, 2-22 (2012). 37, 46, 50

[89] Alonso-Blanco, C., Aarts, M., Bentsink, L., Keurentjes, J., Reymond, M., Vreugdenhil, D. and Koornneef, M., *What has natural variation taught us about plant development, physiology, and adaptation?*, Plant Cell **21(7)**, 1877-1896 (2009). 37, 39, 46

[90] Meyer, R., Steinfath, M., Lisec, J., Becher, M., Witucka-Wall, H., Torjek, O., Fiehn, O., Eckardt, A., Willmitzer, L., Selbig, J. and Altmann, T., *The metabolic signature related to high plant growth rate in Arabidopsis thaliana.*, Proc Natl Acad Sci U S A **104(11)**, 4759-4764 (2007). 37

[91] Sulpice, R., Pyl, E., Ishihara, H., Trenkamp, S., Steinfath, M., Witucka-Wall, H., Gibon, Y., Usadel, B., Poree, F., Piques, M., Von Korff, M., Steinhauser, M., Keurentjes, J., Guenther, M., Hoehne, M., Selbig, J., Fernie, A., Altmann, T. and Stitt, M., *Starch as a major integrator in the regulation of plant growth.*, Proc Natl Acad Sci U S A **106(25)**, 10348-10353 (2009). 37, 193

[92] Prinzenberg, A., Barbier, H., Salt, D., Stich, B. and Reymond, M., *Relationships between growth, growth response to nutrient supply, and ion content using a recombinant inbred line population in Arabidopsis.*, Plant Physiol **154(3)**, 1361-1371 (2010). 37, 57, 60, 86

[93] Kilian, J., Whitehead, D., Horak, J., Wanke, D., Weinl, S., Batistic, O., D'Angelo, C., Bornberg-Bauer, E., Kudla, J. and Harter, K., *The AtGenExpress global stress expression data set: Protocols, evaluation and model data analysis of UV-B light, drought and cold stress responses.*, Plant J **50(2)**, 347-363 (2007). 37

[94] Des Marais, D.L., Mckay, J., Richards, J., Sen, S., Wayne, T. and Juenger, T., *Physiological genomics of response to soil drying in diverse Arabidopsis accessions.*, Plant Cell **24(3)**, 893-914 (2012). 37

[95] Mewis, I., Khan, M., Glawischnig, E., Schreiner, M. and Ulrichs, C., *Water stress and aphid feeding differentially influence metabolite composition in Arabidopsis thaliana (l.).*, PLoS One **7(11)**, e48661 (2012). 37

[96] Sperdouli, I. and Moustakas, M., *Interaction of proline, sugars, and anthocyanins during photosynthetic acclimation of Arabidopsis thaliana to drought stress.*, J Plant Physiol **169(6)**, 577-585 (2012). 37

[97] Ghandilyan, A., Barboza, L., Tisne, S., Granier, C., Reymond, M., Koornneef, M., Schat, H. and Aarts, M., *Genetic analysis identifies quantitative trait loci controlling rosette mineral concentrations in Arabidopsis thaliana under drought.*, New Phytol **184(1)**, 180-192 (2009). 37, 56, 57, 60

[98] Verslues, P. and Juenger, T., *Drought, metabolites, and Arabidopsis natural variation: A promising combination for understanding adaptation to water-limited environments.*, Curr Opin Plant Biol **14(3)**, 240-245 (2011). 37, 39

[99] Brachi, B., Aime, C., Glorieux, C., Cuguen, J. and Roux, F., *Adaptive value of phenological traits in stressful environments: Predictions based on seed production and laboratory natural selection.*, PLoS One **7(3)**, e32069 (2012). 39

[100] Van Kleunen, M. and Fischer, M., *Constraints on the evolution of adaptive phenotypic plasticity in plants.*, New Phytol **166(1)**, 49-60 (2005). 39

[101] Pigliucci, M., *Ecology and evolutionary biology of Arabidopsis.*, Arabidopsis Book **1**, e0003 (2002). 39

[102] Keurentjes, J., Koornneef, M. and Vreugdenhil, D., *Quantitative genetics in the age of omics.*, Curr Opin Plant Biol **11(2)**, 123-128 (2008). 39

[103] Murren, C., *The integrated phenotype.*, Integr Comp Biol **52(1)**, 64-76 (2012). 39

[104] Reiter, R., Williams, J., Feldmann, K., Rafalski, J., Tingey, S. and Scolnik, P., *Global and local genome mapping in Arabidopsis thaliana by using recombinant inbred lines and random amplified polymorphic dnas.*, Proc Natl Acad Sci U S A **89(4)**, 1477-1481 (1992). 44

[105] Lister, C. and Dean, C., *Recombinant inbred lines for mapping RFLP and phenotypic markers in Arabidopsis thaliana*, Plant J **4(4)**, 745-750 (1993). 44

[106] Torjek, O., Witucka-Wall, H., Meyer, R., Von Korff, M., Kusterer, B., Rautengarten, C. and Altmann, T., *Segregation distortion in Arabidopsis C24/Col-0 and Col-0/C24 recombinant inbred line populations is due to reduced fertility caused by epistatic interaction of two loci.*, Theor Appl Genet **113(8)**, 1551-1561 (2006). 44

[107] Simon, M., Loudet, O., Durand, S., Berard, A., Brunel, D., Sennesal, F., Durand-Tardif, M., Pelletier, G. and Camilleri, C., *Quantitative trait loci mapping in five new large recombinant inbred line populations of Arabidopsis thaliana genotyped with consensus single-nucleotide polymorphism markers.*, Genetics **178(4)**, 2253-2264 (2008). 44, 52

[108] Ravi, M. and Chan, S., *Haploid plants produced by centromere-mediated genome elimination.*, Nature **464(7288)**, 615-618 (2010). 44

[109] Seymour, D., Filiault, D., Henry, I., Monson-Miller, J., Ravi, M., Pang, A., Comai, L., Chan, S. and Maloof, J., *Rapid creation of Arabidopsis doubled haploid lines for quantitative trait locus mapping.*, Proc Natl Acad Sci U S A **109(11)**, 4227-4232 (2012). 44

[110] Balasubramanian, S., Schwartz, C., Singh, A., Warthmann, N., Kim, M., Maloof, J., Loudet, O., Trainer, G., Dabi, T., Borevitz, J., Chory, J. and Weigel, D., *QTL mapping in new Arabidopsis thaliana advanced intercross-recombinant inbred lines.*, PLoS One **4(2)**, e4318 (2009). 44

[111] Huang, X., Paulo, M., Boer, M., Effgen, S., Keizer, P., Koornneef, M. and Van Eeuwijk, F.A., *Analysis of natural allelic variation in Arabidopsis using a multiparent recombinant inbred line population.*, Proc Natl Acad Sci U S A **108(11)**, 4488-4493 (2011). 44, 52

[112] Kover, P., Valdar, W., Trakalo, J., Scarcelli, N., Ehrenreich, I., Purugganan, M., Durrant, C. and Mott, R., *A multiparent advanced generation inter-cross to fine-map quantitative traits in Arabidopsis thaliana.*, PLoS Genet **5(7)**, e1000551 (2009). 44, 52

[113] Johannes, F., Porcher, E., Teixeira, F., Saliba-Colombani, V., Simon, M., Agier, N., Bulski, A., Albuisson, J., Heredia, F., Audigier, P., Bouchez, D., Dillmann, C., Guerche, P., Hospital, F. and Colot, V., *Assessing the impact of transgenerational epigenetic variation on complex traits.*, PLoS Genet **5(6)**, e1000530 (2009). 45

[114] Reinders, J., Wulff, B., Mirouze, M., Mari-Ordonez, A., Dapp, M., Rozhon, W., Bucher, E., Theiler, G. and Paszkowski, J., *Compromised stability of DNA methylation and transposon immobilization in mosaic Arabidopsis epigenomes.*, Genes Dev **23(8)**, 939-950 (2009). 45

[115] Roux, F., Colome-Tatche, M., Edelist, C., Wardenaar, R., Guerche, P., Hospital, F., Colot, V., Jansen, R. and Johannes, F., *Genome-wide epigenetic perturbation jump-starts patterns of heritable variation found in nature.*, Genetics **188(4)**, 1015-1017 (2011). 45

[116] Latzel, V., Zhang, Y., Karlsson Moritz, K., Fischer, M. and Bossdorf, O., *Epigenetic variation in plant responses to defence hormones.*, Ann Bot **110(7)**, 1423-1428 (2012). 45

[117] Lander, E. and Botstein, D., *Mapping mendelian factors underlying quantitative traits using RFLP linkage maps.*, Genetics **121(1)**, 185-199 (1989). 46

[118] Lander, E. and Botstein, D., *Mapping complex genetic traits in humans: New methods using a complete RFLP linkage map.*, Cold Spring Harb Symp Quant Biol **51 Pt 1**, 49-62 (1986). 46

[119] Broman, K., *Review of statistical methods for QTL mapping in experimental crosses.*, Lab Anim (NY) **30(7)**, 44-52 (2001). 46

[120] Jansen, R., *Interval mapping of multiple quantitative trait loci.*, Genetics **135(1)**, 205-211 (1993). 46

[121] Zeng, Z., *Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci.*, Proc Natl Acad Sci U S A **90(23)**, 10972-10976 (1993). 46

[122] Zeng, Z., *Precision mapping of quantitative trait loci.*, Genetics **136(4)**, 1457-1468 (1994). 46

[123] Broman, K.W. and Sen S., *A guide to QTL mapping with r/QTL*, **Springer**, (2009). 46

[124] Cowen, N., *Multiple linear regression analysis of RFLP data sets used in mapping QTLs*, Development and Application of Molecular Markers to Problems in Plant Genetics , 113-116 (1989). 46

[125] Kao, C., Zeng, Z. and Teasdale, R., *Multiple interval mapping for quantitative trait loci.*, Genetics **152(3)**, 1203-1216 (1999). 46

[126] Zeng, Z., *QTL mapping and the genetic basis of adaptation: Recent developments.*, Genetica **123(1-2)**, 25-37 (2005). 46

[127] Bergelson, J. and Roux, F., *Towards identifying genes underlying ecologically relevant traits in Arabidopsis thaliana.*, Nat Rev Genet **11(12)**, 867-879 (2010). 47, 50

[128] Tuinstra, M., Ejeta, G. and Goldsbrough, P., *Heterogeneous inbred family (HIF) analysis: A method for developing near-isogenic lines that differ at quantitative trait loci*, Theoretical and Applied Genetics **95(5)**, 1005-1011 (1997). 47

[129] Loudet, O., Gaudon, V., Trubuil, A. and Daniel-Vedele, F., *Quantitative trait loci controlling root growth and architecture in Arabidopsis thaliana confirmed by Heterogeneous Inbred Family.*, Theor Appl Genet **110(4)**, 742-753 (2005). 47

[130] Keurentjes, J., Bentsink, L., Alonso-Blanco, C., Hanhart, C., Blankestijn-De Vries, H., Effgen, S., Vreugdenhil, D. and Koornneef, M., *Development of a near-isogenic line population of Arabidopsis thaliana and comparison of mapping power with a recombinant inbred line population.*, Genetics **175(2)**, 891-905 (2007). 47

[131] Torjek, O., Meyer, R., Zehnsdorf, M., Teltow, M., Strompen, G., Witucka-Wall, H., Blacha, A. and Altmann, T., *Construction and analysis of 2 reciprocal Arabidopsis introgression line populations.*, J Hered **99(4)**, 396-406 (2008). 47

[132] Lisec, J., Meyer, R., Steinfath, M., Redestig, H., Becher, M., Witucka-Wall, H., Fiehn, O., Torjek, O., Selbig, J., Altmann, T. and Willmitzer, L., *Identification of metabolic and biomass QTL in Arabidopsis thaliana in a parallel analysis of RIL and IL populations.*, Plant J **53(6)**, 960-972 (2008). 47, 57, 59

[133] Lisec, J., Steinfath, M., Meyer, R., Selbig, J., Melchinger, A., Willmitzer, L. and Altmann, T., *Identification of heterotic metabolite QTL in Arabidopsis thaliana RIL and IL populations.*, Plant J **59(5)**, 777-788 (2009). 47, 57

[134] Joshi, H., Christiansen, K., Fitz, J., Cao, J., Lipzen, A., Martin, J., Smith-Moritz, A.M., Pennacchio, L., Schackwitz, W., Weigel, D. and Heazlewood, J., *1001 proteomes: A functional proteomics portal for the analysis of Arabidopsis thaliana accessions.*, Bioinformatics **28(10)**, 1303-1306 (2012). 49

[135] Laitinen, R., Schneeberger, K., Jelly, N., Ossowski, S. and Weigel, D., *Identification of a spontaneous frame shift mutation in a nonreference Arabidopsis accession using whole genome sequencing.*, Plant Physiol **153(2)**, 652-654 (2010). 49, 54

[136] Usadel, B., Obayashi, T., Mutwil, M., Giorgi, F., Bassel, G., Tanimoto, M., Chow, A., Steinhauser, D., Persson, S. and Provart, N., *Co-expression tools for plant biology: Opportunities for hypothesis generation and caveats.*, Plant Cell Environ **32(12)**, 1633-1651 (2009). 49

[137] Brady, S. and Provart, N., *Web-queryable large-scale data sets for hypothesis generation in plant biology.*, Plant Cell **21(4)**, 1034-1051 (2009). 49

[138] Wee, C. and Dinneny, J., *Tools for high-spatial and temporal-resolution analysis of environmental responses in plants.*, Biotechnol Lett **32(10)**, 1361-1371 (2010). 49

[139] Jimenez-Gomez, J.M., Wallace, A. and Maloof, J., *Network analysis identifies ELF3 as a QTL for the shade avoidance response in Arabidopsis.*, PLoS Genet **6(9)**, (2010). 49

[140] Kim, S., Plagnol, V., Hu, T., Toomajian, C., Clark, R., Ossowski, S., Ecker, J., Weigel, D. and Nordborg, M., *Recombination and linkage disequilibrium in Arabidopsis thaliana.*, Nat Genet **39(9)**, 1151-1155 (2007). 50

[141] Brachi, B., Faure, N., Horton, M., Flahauw, E., Vazquez, A., Nordborg, M., Bergelson, J., Cuguen, J. and Roux, F., *Linkage and association mapping of Arabidopsis thaliana flowering time in nature.*, PLoS Genet **6(5)**, e1000940 (2010). 50, 52, 60, 65, 204

[142] Li, Y., Huang, Y., Bergelson, J., Nordborg, M. and Borevitz, J., *Association mapping of local climate-sensitive quantitative trait loci in Arabidopsis thaliana.*, Proc Natl Acad Sci U S A **107(49)**, 21199-21204 (2010). 50, 60, 65

[143] Baxter, I., Muthukumar, B., Park, H., Buchner, P., Lahner, B., Danku, J., Zhao, K., Lee, J., Hawkesford, M., Guerinot, M. and Salt, D., *Variation in molybdenum content across broadly distributed populations of Arabidopsis thaliana is controlled by a mitochondrial molybdenum transporter (MOT1).*, PLoS Genet **4(2)**, e1000004 (2008). 50, 86

[144] Tomatsu, H., Takano, J., Takahashi, H., Watanabe-Takahashi, A., Shibagaki, N. and Fujiwara, T., *An Arabidopsis thaliana high-affinity molybdate transporter required for efficient uptake of molybdate from soil.*, Proc Natl Acad Sci U S A **104(47)**, 18807-18812 (2007). 50, 86

[145] Baxter, I., Brazelton, J., Yu, D., Huang, Y., Lahner, B., Yakubova, E., Li, Y., Bergelson, J., Borevitz, J., Nordborg, M., Vitek, O. and Salt, D., *A coastal cline in sodium accumulation in Arabidopsis thaliana is driven by natural variation of the sodium transporter AtHKT1;1.*, PLoS Genet **6(11)**, e1001193 (2010). 50, 66, 79

[146] Rus, A., Baxter, I., Muthukumar, B., Gustin, J., Lahner, B., Yakubova, E. and Salt, D., *Natural variants of AtHKT1 enhance Na+ accumulation in two wild populations of Arabidopsis.*, PLoS Genet **2(12)**, e210 (2006). 50, 54

[147] Todesco, M., Balasubramanian, S., Hu, T., Traw, M., Horton, M., Epple, P., Kuhns, C., Sureshkumar, S., Schwartz, C., Lanz, C., Laitinen, R., Huang, Y., Chory, J., Lipka, V., Borevitz, J., Dangl, J., Bergelson, J., Nordborg, M. and Weigel, D., *Natural allelic variation underlying a major fitness trade-off in Arabidopsis thaliana.*, Nature **465(7298)**, 632-636 (2010). 50, 79

[148] Brachi, B., Morris, G. and Borevitz, J., *Genome-wide association studies in plants: The missing heritability is in the field.*, Genome Biol **12(10)**, 232 (2011). 50, 51

[149] Vilhjalmsson, B. and Nordborg, M., *The nature of confounding in genome-wide association studies.*, Nat Rev Genet **14(1)**, 1-2 (2012). 50

[150] Pritchard, J., Stephens, M. and Donnelly, P., *Inference of population structure using multilocus genotype data.*, Genetics **155(2)**, 945-959 (2000). 50

[151] Myles, S., Peiffer, J., Brown, P., Ersoz, E., Zhang, Z., Costich, D. and Buckler, E., *Association mapping: Critical considerations shift from genotyping to experimental design.*, Plant Cell **21(8)**, 2194-2202 (2009). 51

[152] McMullen, M., Kresovich, S., Villeda, H., Bradbury, P., Li, H., Sun, Q., Flint-Garcia, S., Thornsberry, J., Acharya, C., Bottoms, C., Brown, P., Browne, C., Eller, M., Guill, K., Harjes, C., Kroon, D., Lepak, N., Mitchell, S., Peterson, B., Pressoir, G., Romero, S., Oropeza Rosas, M., Salvo, S., Yates, H., Hanson, M., Jones, E., Smith, S., Glaubitz, J., Goodman, M., Ware, D., Holland, J. and Buckler, E., *Genetic properties of the maize nested association mapping population.*, Science **325(5941)**, 737-740 (2009). 51

[153] Buckler, E., Holland, J., Bradbury, P., Acharya, C., Brown, P., Browne, C., Ersoz, E., Flint-Garcia, S., Garcia, A., Glaubitz, J., Goodman, M., Harjes, C., Guill, K., Kroon, D., Larsson, S., Lepak, N., Li, H., Mitchell, S., Pressoir, G., Peiffer, J., Rosas, M., Rocheford, T., Romay, M., Romero, S., Salvo, S., Sanchez Villeda, H., Da Silva, H.S., Sun, Q., Tian, F., Upadyayula, N., Ware, D., Yates, H., Yu, J., Zhang, Z., Kresovich, S. and Mcmullen, M., *The genetic architecture of maize flowering time.*, Science **325(5941)**, 714-718 (2009). 52

[154] Kump, K., Bradbury, P., Wisser, R., Buckler, E., Belcher, A., Oropeza-Rosas, M.A., Zwonitzer, J., Kresovich, S., Mcmullen, M., Ware, D., Balint-Kurti, P.J. and Holland, J., *Genome-wide*

association study of quantitative resistance to southern leaf blight in the maize nested association mapping population., Nat Genet **43(2)**, 163-168 (2011). 52

[155] Tian, F., Bradbury, P., Brown, P., Hung, H., Sun, Q., Flint-Garcia, S., Rocheford, T., Mcmullen, M., Holland, J. and Buckler, E., *Genome-wide association study of leaf architecture in the maize nested association mapping population.*, Nat Genet **43(2)**, 159-162 (2011). 52

[156] Bentsink, L., Hanson, J., Hanhart, C., Vries, H., Coltrane, C., Keizer, P., El-Lithy, M., Alonso-Blanco, C., Teresa De,res, M., Reymond, M., Van Eeuwijk, F., Smeekens, S. and Koornneef, M., *Natural variation for seed dormancy in Arabidopsis is regulated by additive genetic and molecular pathways.*, Proc Natl Acad Sci U S A **107(9)**, 4264-4269 (2010). 52

[157] Ossowski, S., Schwab, R. and Weigel, D., *Gene silencing in plants using artificial micrornas and other small RNAs.*, Plant J **53(4)**, 674-690 (2008). 52

[158] Mackay, T., *The genetic architecture of quantitative traits.*, Annu Rev Genet **35**, 303-339 (2001). 52, 56

[159] Schwartz, C., Balasubramanian, S., Warthmann, N., Michael, T., Lempe, J., Sureshkumar, S., Kobayashi, Y., Maloof, J., Borevitz, J., Chory, J. and Weigel, D., *Cis-regulatory changes at FLOWERING LOCUS T mediate natural variation in flowering responses of Arabidopsis thaliana.*, Genetics **183(2)**, 723-32, 1SI-7SI (2009). 54

[160] Loudet, O., Saliba-Colombani, V., Camilleri, C., Calenge, F., Gaudon, V., Koprivova, A., North, K., Kopriva, S. and Daniel-Vedele, F., *Natural variation for sulfate content in Arabidopsis thaliana is highly controlled by APR2.*, Nat Genet **39(7)**, 896-900 (2007). 54, 56, 61, 70

[161] Ren, Z., Zheng, Z., Chinnusamy, V., Zhu, J., Cui, X., Iida, K. and Zhu, J., *RAS1, a quantitative trait locus for salt tolerance and aba sensitivity in Arabidopsis.*, Proc Natl Acad Sci U S A **107(12)**, 5669-5674 (2010). 54, 56, 70

[162] Kesari, R., Lasky, J., Villamor, J., Des Marais, D.L., Chen, Y., Liu, T., Lin, W., Juenger, T. and Verslues, P., *Intron-mediated alternative splicing of Arabidopsis P5CS1 and its association with natural variation in proline and climate adaptation.*, Proc Natl Acad Sci U S A **109(23)**, 9197-9202 (2012). 54

[163] Svistoonoff, S., Creff, A., Reymond, M., Sigoillot-Claude, C., Ricaud, L., Blanchet, A., Nussaume, L. and Desnos, T., *Root tip contact with low-phosphate media reprograms plant root architecture.*, Nat Genet **39(6)**, 792-796 (2007). 54

[164] Nemri, A., Atwell, S., Tarone, A., Huang, Y., Zhao, K., Studholme, D., Nordborg, M. and Jones, J., *Genome-wide survey of Arabidopsis natural variation in downy mildew resistance using combined association and linkage mapping.*, Proc Natl Acad Sci U S A **107(22)**, 10302-10307 (2010). 56

[165] Ghandilyan, A., Ilk, N., Hanhart, C., Mbengue, M., Barboza, L., Schat, H., Koornneef, M., El-Lithy, M., Vreugdenhil, D., Reymond, M. and Aarts, M., *A strong effect of growth medium and organ type on the identification of QTLs for phytate and mineral concentrations in three Arabidopsis thaliana ril populations.*, J Exp Bot **60(5)**, 1409-1425 (2009). 56, 60

[166] Buescher, E., Achberger, T., Amusan, I., Giannini, A., Ochsenfeld, C., Rus, A., Lahner, B., Hoekenga, O., Yakubova, E., Harper, J., Guerinot, M., Zhang, M., Salt, D. and Baxter, I., *Natural genetic variation in selected populations of Arabidopsis thaliana is associated with ionomic differences.*, PLoS One **5(6)**, e11081 (2010). 56, 57, 60, 86, 121

[167] Segura, V., Vilhjalmsson, B., Platt, A., Korte, A., Seren, U., Long, Q. and Nordborg, M., *An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations.*, Nat Genet **44(7)**, 825-830 (2012). 56, 57

[168] Le Corre, V., Roux, F. and Reboud, X., *DNA polymorphism at the FRIGIDA gene in Arabidopsis thaliana: Extensive nonsynonymous variation is consistent with local selection for flowering time.*, Mol Biol Evol **19(8)**, 1261-1271 (2002). 56, 75

[169] Shindo, C., Aranzana, M., Lister, C., Baxter, C., Nicholls, C., Nordborg, M. and Dean, C., *Role of FRIGIDA and FLOWERING LOCUS C in determining variation in flowering time of Arabidopsis.*, Plant Physiol **138(2)**, 1163-1173 (2005). 56

[170] Lempe, J., Balasubramanian, S., Sureshkumar, S., Singh, A., Schmid, M. and Weigel, D., *Diversity of flowering responses in wild Arabidopsis thaliana strains.*, PLoS Genet **1(1)**, 109-118 (2005). 56

[171] Werner, J., Borevitz, J., Warthmann, N., Trainer, G., Ecker, J., Chory, J. and Weigel, D., *Quantitative trait locus mapping and DNA array hybridization identify an flm deletion as a cause for natural flowering-time variation.*, Proc Natl Acad Sci U S A **102(7)**, 2460-2465 (2005). 56

[172] Michaels, S., He, Y., Scortecci, K. and Amasino, R., *Attenuation of FLOWERING LOCUS C activity as a mechanism for the evolution of summer-annual flowering behavior in Arabidopsis.*, Proc Natl Acad Sci U S A **100(17)**, 10102-10107 (2003). 56

[173] Mendez-Vigo, B., Pico, F., Ramiro, M., Martinez-Zapater, J.M. and Alonso-Blanco, C., *Altitudinal and climatic adaptation is mediated by flowering traits and FRI, FLC, and PHYC genes in Arabidopsis.*, Plant Physiol **157(4)**, 1942-1955 (2011). 56, 65, 66, 78

[174] Sanchez-Bermejo, E., Mendez-Vigo, B., Pico, F., Martinez-Zapater, J.M. and Alonso-Blanco, C., *Novel natural alleles at FLC and LVR loci account for enhanced vernalization responses in Arabidopsis thaliana.*, Plant Cell Environ **35(9)**, 1672-1684 (2012). 56

[175] Kobayashi, Y., Kuroda, K., Kimura, K., Southron-Francis, J.L., Furuzawa, A., Kimura, K., Iuchi, S., Kobayashi, M., Taylor, G. and Koyama, H., *Amino acid polymorphisms in strictly conserved domains of a P-type ATPase HMA5 are involved in the mechanism of copper tolerance variation in Arabidopsis.*, Plant Physiol **148(2)**, 969-980 (2008). 57

[176] Chao, D., Silva, A., Baxter, I., Huang, Y., Nordborg, M., Danku, J., Lahner, B., Yakubova, E. and Salt, D., *Genome-wide association studies identify heavy metal ATPase3 as the primary determinant of natural variation in leaf cadmium in Arabidopsis thaliana.*, PLoS Genet **8(9)**, e1002923 (2012). 57

[177] Fu, J., Keurentjes, J., Bouwmeester, H., America, T., Verstappen, F., Ward, J., Beale, M., De Vos, R.C., Dijkstra, M., Scheltema, R., Johannes, F., Koornneef, M., Vreugdenhil, D., Breitling, R. and Jansen, R., *System-wide molecular evidence for phenotypic buffering in Arabidopsis.*, Nat Genet **41(2)**, 166-167 (2009). 57, 59, 60, 204

[178] Keurentjes, J., Sulpice, R., Gibon, Y., Steinhauser, M., Fu, J., Koornneef, M., Stitt, M. and Vreugdenhil, D., *Integrative analyses of genetic variation in enzyme activities of primary carbohydrate metabolism reveal distinct modes of regulation in Arabidopsis thaliana.*, Genome Biol **9(8)**, R129 (2008). 57, 59, 60

[179] Rowe, H., Hansen, B., Halkier, B. and Kliebenstein, D., *Biochemical networks and epistasis shape the Arabidopsis thaliana metabolome.*, Plant Cell **20(5)**, 1199-1216 (2008). 57, 59, 60

[180] Keurentjes, J., Fu, J., Terpstra, I., Garcia, J., Van Den Ackerveken, G., Snoek, L., Peeters, A., Vreugdenhil, D., Koornneef, M. and Jansen, R., *Regulatory network construction in Arabidopsis by using genome-wide gene expression quantitative trait loci.*, Proc Natl Acad Sci U S A **104(5)**, 1708-1713 (2007). 57, 203

[181] Cubillos, F., Yansouni, J., Khalili, H., Balzergue, S., Elftieh, S., Martin-Magniette, M.L., Serrand, Y., Lepiniec, L., Baud, S., Dubreucq, B., Renou, J., Camilleri, C. and Loudet, O., *Expression variation in connected recombinant populations of Arabidopsis thaliana highlights distinct transcriptome architectures.*, BMC Genomics **13**, 117 (2012). 57, 173, 203

[182] Tisne, S., Reymond, M., Vile, D., Fabre, J., Dauzat, M., Koornneef, M. and Granier, C., *Combined genetic and modeling approaches reveal that epidermal cell area and number in leaves are controlled by leaf and plant developmental processes in Arabidopsis.*, Plant Physiol **148(2)**, 1117-1127 (2008). 57

[183] Gardner, K. and Latta, R., *Shared quantitative trait loci underlying the genetic correlation between continuous traits.*, Mol Ecol **16(20)**, 4195-4209 (2007). 57

[184] Carreno-Quintero, N., Bouwmeester, H. and Keurentjes, J., *Genetic analysis of metabolome-phenotype interactions: From model to crop species.*, Trends Genet **29(1)**, 41-50 (2013). 58

[185] Chan, E., Rowe, H. and Kliebenstein, D., *Understanding the evolution of defense metabolites in Arabidopsis thaliana using genome-wide association mapping.*, Genetics **185(3)**, 991-1007 (2010). 58

[186] Van Eeuwijk, F.A., Bink, M., Chenu, K. and Chapman, S., *Detection and use of QTL for complex traits in multiple environments.*, Curr Opin Plant Biol **13(2)**, 193-205 (2010). 58, 60

[187] Korte, A., Vilhjalmsson, B., Segura, V., Platt, A., Long, Q. and Nordborg, M., *A mixed-model approach for genome-wide association studies of correlated traits in structured populations.*, Nat Genet **44(9)**, 1066-1071 (2012). 58

[188] Phillips, P., *Epistasis–the essential role of gene interactions in the structure and evolution of genetic systems.*, Nat Rev Genet **9(11)**, 855-867 (2008). 58

[189] Kroymann, J. and Mitchell-Olds, T., *Epistasis and balanced polymorphism influencing complex trait variation.*, Nature **435(7038)**, 95-98 (2005). 58

[190] Denby, K., Kumar, P. and Kliebenstein, D., *Identification of Botrytis cinerea susceptibility loci in Arabidopsis thaliana.*, Plant J **38(3)**, 473-486 (2004). 59

[191] Zhang, Z., Ober, J. and Kliebenstein, D., *The gene controlling the quantitative trait locus EPITHIOSPECIFIER MODIFIER1 alters glucosinolate hydrolysis and insect resistance in Arabidopsis.*, Plant Cell **18(6)**, 1524-1536 (2006). 59

[192] Malmberg, R., Held, S., Waits, A. and Mauricio, R., *Epistasis for fitness-related quantitative traits in Arabidopsis thaliana grown in the field and in the greenhouse.*, Genetics **171(4)**, 2013-2027 (2005). 59, 65

[193] Juenger, T., Sen, S., Stowe, K. and Simms, E., *Epistasis and genotype-environment interaction for quantitative trait loci affecting flowering time in Arabidopsis thaliana.*, Genetica **123(1-2)**, 87-105 (2005). 59

[194] Brock, M., Tiffin, P. and Weinig, C., *Sequence diversity and haplotype associations with phenotypic responses to crowding: GIGANTEA affects fruit set in Arabidopsis thaliana.*, Mol Ecol **16(14)**, 3050-3062 (2007). 59

[195] Loudet, O., Chaillou, S., Merigout, P., Talbotec, J. and Daniel-Vedele, F., *Quantitative trait loci analysis of nitrogen use efficiency in Arabidopsis.*, Plant Physiol **131(1)**, 345-358 (2003). 59

[196] Calenge, F., Saliba-Colombani, V., Mahieu, S., Loudet, O., Daniel-Vedele, F. and Krapp, A., *Natural variation for carbohydrate content in Arabidopsis. Interaction with complex traits dissected by quantitative genetics.*, Plant Physiol **141(4)**, 1630-1643 (2006). 59

[197] Alcazar, R., Garcia, A., Parker, J. and Reymond, M., *Incremental steps toward incompatibility revealed by Arabidopsis epistatic interactions modulating salicylic acid pathway activation.*, Proc Natl Acad Sci U S A **106(1)**, 334-339 (2009). 59

[198] Alcazar, R., Garcia, A., Kronholm, I., De Meaux, J., Koornneef, M., Parker, J. and Reymond, M., *Natural variation at strubbelig receptor kinase 3 drives immune-triggered incompatibilities between Arabidopsis thaliana accessions.*, Nat Genet **42(12)**, 1135-1139 (2010). 59

[199] Bikard, D., Patel, D., Le Mette, C., Giorgi, V., Camilleri, C., Bennett, M. and Loudet, O., *Divergent evolution of duplicate genes leads to genetic incompatibilities within A. thaliana.*, Science **323(5914)**, 623-626 (2009). 59

[200] Ehrenreich, I., Stafford, P. and Purugganan, M., *The genetic architecture of shoot branching in Arabidopsis thaliana: A comparative assessment of candidate gene associations vs. Quantitative trait locus mapping.*, Genetics **176(2)**, 1223-1236 (2007). 59

[201] Wentzell, A., Rowe, H., Hansen, B., Ticconi, C., Halkier, B. and Kliebenstein, D., *Linking metabolic QTLs with network and cis-eQTLs controlling biosynthetic pathways.*, PLoS Genet **3(9)**, 1687-1701 (2007). 59

[202] Jansen, R., Tesson, B., Fu, J., Yang, Y. and Mcintyre, L., *Defining gene and QTL networks.*, Curr Opin Plant Biol **12(2)**, 241-246 (2009). 60

[203] Snoek, L., Terpstra, I., Dekter, R., Van Den Ackerveken, G. and Peeters, A., *Genetical genomics reveals large scale genotype-by-environment interactions in Arabidopsis thaliana.*, Front Genet **3**, 317 (2012). 60

[204] Chan, E., Rowe, H., Corwin, J., Joseph, B. and Kliebenstein, D., *Combining genome-wide association mapping and transcriptional networks to identify novel genes controlling glucosinolates in Arabidopsis thaliana.*, PLoS Biol **9(8)**, e1001125 (2011). 60

[205] Fournier-Level, A., Wilczek, A., Cooper, M., Roe, J., Anderson, J., Eaton, D., Moyers, B., Petipas, R., Schaeffer, R., Pieper, B., Reymond, M., Koornneef, M., Welch, S., Remington, D. and Schmitt, J., *Paths to selection on life history loci in different natural environments across the native range of Arabidopsis thaliana.*, Mol Ecol, mec.12285 (2013). 60, 61, 67

[206] Li, Y., Roycewicz, P., Smith, E. and Borevitz, J., *Genetics of local adaptation in the laboratory: Flowering time quantitative trait loci under geographic and seasonal conditions in Arabidopsis.*, PLoS One **1**, e105 (2006). 60, 65

[207] Matesanz, S., Gianoli, E. and Valladares, F., *Global change and the evolution of phenotypic plasticity in plants.*, Ann N Y Acad Sci **1206**, 35-55 (2010). 61, 67

[208] Pigliucci, M., *Evolution of phenotypic plasticity: Where are we going now?*, Trends Ecol Evol **20(9)**, 481-486 (2005). 61

[209] Sibout, R., Plantegenet, S. and Hardtke, C., *Flowering as a condition for xylem expansion in Arabidopsis hypocotyl and root.*, Curr Biol **18(6)**, 458-463 (2008). 62

[210] Orr, H., *Fitness and its role in evolutionary genetics.*, Nat Rev Genet **10(8)**, 531-539 (2009). 63, 65, 67

[211] Fournier-Level, A., Korte, A., Cooper, M., Nordborg, M., Schmitt, J. and Wilczek, A., *A map of local adaptation in Arabidopsis thaliana.*, Science **334(6052)**, 86-89 (2011). 64, 65, 66, 67, 76, 79, 204

[212] Huang, X., Schmitt, J., Dorn, L., Griffith, C., Effgen, S., Takao, S., Koornneef, M. and Donohue, K., *The earliest stages of adaptation in an experimental plant population: Strong selection on QTLs for seed dormancy.*, Mol Ecol **19(7)**, 1335-1351 (2010). 64, 79

[213] De Jong, G., *The fitness of fitness concepts and the description of natural selection*, The Quarterly Review of Biology **69(1)**, 3-29 (1994). 64

[214] Donohue, K., Dorn, L., Griffith, C., Kim, E., Aguilera, A., Polisetty, C. and Schmitt, J., *The evolutionary ecology of seed germination of Arabidopsis thaliana: Variable natural selection on germination timing.*, Evolution **59(4)**, 758-770 (2005). 64, 66

[215] Korves, T., Schmid, K., Caicedo, A., Mays, C., Stinchcombe, J., Purugganan, M. and Schmitt, J., *Fitness effects associated with the major flowering time gene FRIGIDA in Arabidopsis thaliana in the field.*, Am Nat **169(5)**, E141-57 (2007). 64

[216] Wilczek, A., Roe, J., Knapp, M., Cooper, M., Lopez-Gallego, C., Martin, L., Muir, C., Sim, S., Walker, A., Anderson, J., Egan, J., Moyers, B., Petipas, R., Giakountis, A., Charbit, E.,

Coupland, G., Welch, S. and Schmitt, J., *Effects of genetic perturbation on seasonal life history plasticity.*, Science **323(5916)**, 930-934 (2009). 64

[217] Weinig, C., Ungerer, M., Dorn, L., Kane, N., Toyonaga, Y., Halldorsdottir, S., Mackay, T., Purugganan, M. and Schmitt, J., *Novel loci control variation in reproductive timing in Arabidopsis thaliana in natural environments.*, Genetics **162(4)**, 1875-1884 (2002). 65, 66

[218] Massonnet, C., Vile, D., Fabre, J., Hannah, M., Caldana, C., Lisec, J., Beemster, G., Meyer, R., Messerli, G., Gronlund, J., Perkovic, J., Wigmore, E., May, S., Bevan, M., Meyer, C., Rubio-Diaz, S., Weigel, D., Micol, J., Buchanan-Wollaston, V., Fiorani, F., Walsh, S., Rinn, B., Gruissem, W., Hilson, P., Hennig, L., Willmitzer, L. and Granier, C., *Probing the reproducibility of leaf growth and molecular phenotypes: A comparison of three Arabidopsis accessions cultivated in ten laboratories.*, Plant Physiol **153(4)**, 2142-2157 (2010). 65

[219] Maloof, J., Borevitz, J., Dabi, T., Lutes, J., Nehring, R., Redfern, J., Trainer, G., Wilson, J., Asami, T., Berry, C., Weigel, D. and Chory, J., *Natural variation in light sensitivity of Arabidopsis.*, Nat Genet **29(4)**, 441-446 (2001). 65

[220] Samis, K., Heath, K. and Stinchcombe, J., *Discordant longitudinal clines in flowering time and Phytochrome C in Arabidopsis thaliana.*, Evolution **62(12)**, 2971-2983 (2008). 65, 66

[221] Samis, K., Murren, C., Bossdorf, O., Donohue, K., Fenster, C., Malmberg, R., Purugganan, M. and Stinchcombe, J., *Longitudinal trends in climate drive flowering time clines in north american Arabidopsis thaliana.*, Ecol Evol **2(6)**, 1162-1180 (2012). 65, 66

[222] Balasubramanian, S., Sureshkumar, S., Agrawal, M., Michael, T., Wessinger, C., Maloof, J., Clark, R., Warthmann, N., Chory, J. and Weigel, D., *The PHYTOCHROME C photoreceptor gene mediates natural variation in flowering and growth responses of Arabidopsis thaliana.*, Nat Genet **38(6)**, 711-715 (2006). 65

[223] Stinchcombe, J., Weinig, C., Ungerer, M., Olsen, K., Mays, C., Halldorsdottir, S., Purugganan, M. and Schmitt, J., *A latitudinal cline in flowering time in Arabidopsis thaliana modulated by the flowering time gene FRIGIDA.*, Proc Natl Acad Sci U S A **101(13)**, 4712-4717 (2004). 66

[224] Chew, Y., Wilczek, A., Williams, M., Welch, S., Schmitt, J. and Halliday, K., *An augmented Arabidopsis phenology model reveals seasonal temperature control of flowering time.*, New Phytol **194(3)**, 654-665 (2012). 66, 78

[225] Banta, J., Ehrenreich, I., Gerard, S., Chou, L., Wilczek, A., Schmitt, J., Kover, P. and Purugganan, M., *Climate envelope modelling reveals intraspecific relationships among flowering phenology, niche breadth and potential range size in Arabidopsis thaliana.*, Ecol Lett **15(8)**, 769-777 (2012). 66, 78

[226] Montesinos-Navarro, A., Pico, F. and Tonsor, S., *Clinal variation in seed traits influencing life cycle timing in Arabidopsis thaliana.*, Evolution **66(11)**, 3417-3431 (2012). 66, 78

[227] Lev-Yadun, S. and Berleth, T., *Expanding ecological and evolutionary insights from wild Arabidopsis thaliana accessions.*, Plant Signal Behav **4(8)**, 796-797 (2009). 66

[228] Trontin, C., Tisne, S., Bach, L. and Loudet, O., *What does Arabidopsis natural variation teach us (and does not teach us) about adaptation in plants?*, Curr Opin Plant Biol **14(3)**, 225-231 (2011). 66

[229] Ungerer, M. and Rieseberg, L., *Genetic architecture of a selection response in Arabidopsis thaliana.*, Evolution **57(11)**, 2531-2539 (2003). 66

[230] Ungerer, M., Linder, C. and Rieseberg, L., *Effects of genetic background on response to selection in experimental populations of Arabidopsis thaliana.*, Genetics **163(1)**, 277-286 (2003). 66

[231] Scarcelli, N. and Kover, P., *Standing genetic variation in FRIGIDA mediates experimental evolution of flowering time in Arabidopsis.*, Mol Ecol **18(9)**, 2039-2049 (2009). 66

[232] Fakheran, S., Paul-Victor, C., Heichinger, C., Schmid, B., Grossniklaus, U. and Turnbull, L., *Adaptation and extinction in experimentally fragmented landscapes.*, Proc Natl Acad Sci U S A **107(44)**, 19120-19125 (2010). 66

[233] Weinig, C., Dorn, L., Kane, N., German, Z., Halldorsdottir, S., Ungerer, M., Toyonaga, Y., Mackay, T., Purugganan, M. and Schmitt, J., *Heterogeneous selection at specific loci in natural environments in Arabidopsis thaliana.*, Genetics **165(1)**, 321-329 (2003). 66

[234] Donohue, K., Dorn, L., Griffith, C., Kim, E., Aguilera, A., Polisetty, C. and Schmitt, J., *Niche construction through germination cueing: Life-history responses to timing of germination in Arabidopsis thaliana.*, Evolution **59(4)**, 771-785 (2005). 66

[235] Agren, J. and Schemske, D., *Reciprocal transplants demonstrate strong adaptive differentiation of the model organism Arabidopsis thaliana in its native range.*, New Phytol **194(4)**, 1112-1122 (2012). 66, 67

[236] Colautti, R., Lee, C. and Mitchell-Olds, T., *Origin, fate, and architecture of ecologically relevant genetic variation.*, Curr Opin Plant Biol **15(2)**, 199-204 (2012). 67

[237] Pagan, I., Fraile, A., Fernandez-Fueyo, E., Montes, N., Alonso-Blanco, C. and Garcia-Arenal, F., *Arabidopsis thaliana as a model for the study of plant-virus co-evolution.*, Philos Trans R Soc Lond B Biol Sci **365(1548)**, 1983-1995 (2010). 68

[238] Wright, S. and Gaut, B., *Molecular population genetics and the search for adaptive evolution in plants.*, Mol Biol Evol **22(3)(3)**, 506-519 (2005). 69, 70

[239] Nielsen, R., *Molecular signatures of natural selection.*, Annu Rev Genet **39**, 197-218 (2005). 69, 70, 76

[240] Mitchell-Olds, T., Willis, J. and Goldstein, D., *Which evolutionary processes influence natural genetic variation for phenotypic traits?*, Nat Rev Genet **8(11)**, 845-856 (2007). 69, 70, 79

[241] Suzuki, Y., *Statistical methods for detecting natural selection from genomic data.*, Genes Genet Syst **85(6)**, 359-376 (2010). 69, 75, 76, 77

[242] Walsh, B. and Lynch, M., *Volume 2: Evolution and selection of quantitative traits*, Online pre-publication version, 2013 . 69, 75, 76, 77

[243] Nordborg, M., Borevitz, J., Bergelson, J., Berry, C., Chory, J., Hagenblad, J., Kreitman, M., Maloof, J., Noyes, T., Oefner, P., Stahl, E. and Weigel, D., *The extent of linkage disequilibrium in Arabidopsis thaliana.*, Nat Genet **30(2)**, 190-193 (2002). 70, 74

[244] Tajima, F., *Statistical method for testing the neutral mutation hypothesis by DNA polymorphism.*, Genetics **123(3)**, 585-595 (1989). 74, 75

[245] Watterson, G., *On the number of segregating sites in genetical models without recombination.*, Theor Popul Biol **7(2)**, 256-276 (1975). 74

[246] Fu, Y. and Li, W., *Statistical tests of neutrality of mutations.*, Genetics **133(3)**, 693-709 (1993). 74, 75

[247] Fay, J. and Wu, C., *Hitchhiking under positive darwinian selection.*, Genetics **155(3)**, 1405-1413 (2000). 74, 75

[248] Marjoram, P. and Tavare, S., *Modern computational approaches for analysing molecular genetic variation data.*, Nat Rev Genet **7(10)**, 759-770 (2006). 74

[249] Excoffier, L. and Ray, N., *Surfing during population expansions promotes genetic revolutions and structuration.*, Trends Ecol Evol **23(7)**, 347-351 (2008). 75, 114

[250] Hancock, A., Brachi, B., Faure, N., Horton, M., Jarymowycz, L., Sperone, F., Toomajian, C., Roux, F. and Bergelson, J., *Adaptation to climate across the Arabidopsis thaliana genome.*, Science **334(6052)**, 83-86 (2011). 75, 76, 78, 79

[251] Toomajian, C., Hu, T., Aranzana, M., Lister, C., Tang, C., Zheng, H., Zhao, K., Calabrese, P., Dean, C. and Nordborg, M., *A nonparametric test reveals selection for rapid flowering in the Arabidopsis genome.*, PLoS Biol **4(5)**, e137 (2006). 76

[252] Hudson, R., Kreitman, M. and Aguade, M., *A test of neutral molecular evolution based on nucleotide data.*, Genetics **116(1)**, 153-159 (1987). 77

[253] Mcdonald, J. and Kreitman, M., *Adaptive protein evolution at the ADH locus in drosophila.*, Nature **351(6328)**, 652-654 (1991). 78

[254] Rand, D. and Kann, L., *Excess amino acid polymorphism in mitochondrial DNA: Contrasts among genes from drosophila, mice, and humans.*, Mol Biol Evol **13(6)**, 735-748 (1996). 78

[255] Eyre-Walker, A. and Keightley, P., *Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change.*, Mol Biol Evol **26(9)**, 2097-2108 (2009). 78

[256] Charlesworth, J. and Eyre-Walker, A., *The McDonald-Kreitman test and slightly deleterious mutations.*, Mol Biol Evol **25(6)**, 1007-1015 (2008). 78

[257] Parsch, J., Zhang, Z. and Baines, J., *The influence of demography and weak selection on the McDonald-Kreitman test: An empirical study in drosophila.*, Mol Biol Evol **26(3)**, 691-698 (2009). 78

[258] Eyre-Walker, A., *Changing effective population size and the McDonald-Kreitman test.*, Genetics **162(4)**, 2017-2024 (2002). 78

[259] Zust, T., Heichinger, C., Grossniklaus, U., Harrington, R., Kliebenstein, D. and Turnbull, L., *Natural enemies drive geographic variation in plant defenses.*, Science **338(6103)**, 116-119 (2012). 79

[260] Tian, D., Traw, M., Chen, J., Kreitman, M. and Bergelson, J., *Fitness costs of R-gene-mediated resistance in Arabidopsis thaliana.*, Nature **423(6935)**, 74-77 (2003). 79

[261] Rose, L., Bittner-Eddy, P.D., Langley, C., Holub, E., Michelmore, R. and Beynon, J., *The maintenance of extreme amino acid diversity at the disease resistance gene, RPP13, in Arabidopsis thaliana.*, Genetics **166(3)**, 1517-1527 (2004). 79, 168

[262] Rose, L., Atwell, S., Grant, M. and Holub, E., *Parallel loss-of-function at the RPM1 bacterial resistance locus in Arabidopsis thaliana.*, Front Plant Sci **3**, 287 (2012). 79, 168

[263] Hall, S., Allen, R., Baumber, R., Baxter, L., Fisher, K., Bittner-Eddy, P.D., Rose, L., Holub, E. and Beynon, J., *Maintenance of genetic variation in plants and pathogens involves complex networks of gene-for-gene interactions.*, Mol Plant Pathol **10(4)**, 449-457 (2009). 79, 168

[264] Bakker, E., Toomajian, C., Kreitman, M. and Bergelson, J., *A genome-wide survey of r gene polymorphisms in Arabidopsis.*, Plant Cell **18(8)**, 1803-1818 (2006). 79, 168

[265] Bakker, E., Traw, M., Toomajian, C., Kreitman, M. and Bergelson, J., *Low levels of polymorphism in genes that control the activation of defense response in Arabidopsis thaliana.*, Genetics **178(4)**, 2031-2043 (2008). 79, 168

[266] Kaiser, B., Gridley, K., Ngaire Brady, J., Phillips, T. and Tyerman, S., *The role of molybdenum in agricultural plant production.*, Ann Bot (Lond) **96(5)**, 745-754 (2005). 85, 116

[267] Seo, M., Peeters, A., Koiwai, H., Oritani, T., Marion-Poll, A., Zeevaart, J., Koornneef, M., Kamiya, Y. and Koshiba, T., *The Arabidopsis aldehyde oxidase 3 (AAO3) gene product catalyzes the final step in abscisic acid biosynthesis in leaves.*, Proc Natl Acad Sci U S A **97(23)**, 12908-12913 (2000). 85

[268] Ibdah, M., Chen, Y., Wilkerson, C. and Pichersky, E., *An aldehyde oxidase in developing seeds of Arabidopsis converts benzaldehyde to benzoic acid.*, Plant Physiol **150(1)**, 416-423 (2009). 85

[269] Bittner, F., Oreb, M. and Mendel, R., *ABA3 is a molybdenum cofactor sulfurase required for activation of aldehyde oxidase and xanthine dehydrogenase in Arabidopsis thaliana.*, J Biol Chem **276(44)**, 40381-40384 (2001). 85

[270] Xiong, L., Ishitani, M., Lee, H. and Zhu, J., *The Arabidopsis LOS5/ABA3 locus encodes a molybdenum cofactor sulfurase and modulates cold stress- and osmotic stress-responsive gene expression.*, Plant Cell **13(9)**, 2063-2083 (2001). 85

[271] Nowak, K., Luniak, N., Witt, C., Wustefeld, Y., Wachter, A., Mendel, R. and Hansch, R., *Peroxisomal localization of sulfite oxidase separates it from chloroplast-based sulfur assimilation.*, Plant Cell Physiol **45(12)**, 1889-1894 (2004). 85

[272] Zrenner, R., Stitt, M., Sonnewald, U. and Boldt, R., *Pyrimidine and purine biosynthesis and degradation in plants.*, Annu Rev Plant Biol **57**, 805-836 (2006). 85

[273] Mendel, R., *Biology of the molybdenum cofactor.*, J Exp Bot **58(9)**, 2289-2296 (2007). 85

[274] Mendel, R. and Hansch, R., *Molybdoenzymes and molybdenum cofactor in plants.*, J Exp Bot **53(375)**, 1689-1698 (2002). 85

[275] Self, W., Grunden, A., Hasona, A. and Shanmugam, K., *Molybdate transport.*, Res Microbiol **152(3-4)**, 311-321 (2001). 85

[276] Buchner, P., Takahashi, H. and Hawkesford, M., *Plant sulphate transporters: Co-ordination of uptake, intracellular and long-distance transport.*, J Exp Bot **55(404)**, 1765-1773 (2004). 85

[277] Gasber, A., Klaumann, S., Trentmann, O., Trampczynska, A., Clemens, S., Schneider, S., Sauer, N., Feifer, I., Bittner, F., Mendel, R. and Neuhaus, H., *Identification of an Arabidopsis solute carrier critical for intracellular transport and inter-organ allocation of molybdate.*, Plant Biol (Stuttg) **13(5)**, 710-718 (2011). 86, 109

[278] Terry, N., Zayed, A., De Souza, M.P. and Tarun, A., *Selenium in higher plants.*, Annu Rev Plant Physiol Plant Mol Biol **51**, 401-432 (2000). 116

[279] Shibagaki, N., Rose, A., Mcdermott, J., Fujiwara, T., Hayashi, H., Yoneyama, T. and Davies, J., *Selenate-resistant mutants of Arabidopsis thaliana identify SULTR1;2, a sulfate transporter required for efficient transport of sulfate into roots.*, Plant J **29(4)**, 475-486 (2002). 116

[280] El Kassis, E., Cathala, N., Rouached, H., Fourcroy, P., Berthomieu, P., Terry, N. and Davidian, J., *Characterization of a selenate-resistant Arabidopsis mutant. Root growth as a potential target for selenate toxicity.*, Plant Physiol **143(3)**, 1231-1241 (2007). 116

[281] Barberon, M., Berthomieu, P., Clairotte, M., Shibagaki, N., Davidian, J. and Gosti, F., *Unequal functional redundancy between the two Arabidopsis thaliana high-affinity sulphate transporters SULTR1;1 and SULTR1;2.*, New Phytol **180(3)**, 608-619 (2008). 116

[282] Zhang, L., Byrne, P. and Pilon-Smits, E.A., *Mapping quantitative trait loci associated with selenate tolerance in Arabidopsis thaliana.*, New Phytol **170(1)**, 33-42 (2006). 117

[283] Lowry, D., Sheng, C., Zhu, Z., Juenger, T., Lahner, B., Salt, D. and Willis, J., *Mapping of ionomic traits in mimulus guttatus reveals Mo and Cd QTLs that colocalize with mot1 homologues.*, PLoS One **7(1)**, e30730 (2012). 121

[284] Stoop, J., Williamson, J. and Mason Pharr, D., *Mannitol metabolism in plants: A method for coping with stress*, Trends in Plant Science **1(5)**, 139-144 (1996). 125, 157

[285] Barrett, L., Kniskern, J., Bodenhausen, N., Zhang, W. and Bergelson, J., *Continua of specificity and virulence in plant host-pathogen interactions: Causes and consequences.*, New Phytol **183(3)**, 513-529 (2009). 153

[286] Boller, T. and Felix, G., *A renaissance of elicitors: Perception of microbe-associated molecular patterns and danger signals by pattern-recognition receptors.*, Annu Rev Plant Biol **60**, 379-406 (2009). 153, 154, 161

[287] Tsuda, K. and Katagiri, F., *Comparing signaling mechanisms engaged in pattern-triggered and effector-triggered immunity.*, Curr Opin Plant Biol **13(4)**, 459-465 (2010). 153

[288] Skirycz, A., De Bodt, S., Obata, T., De Clercq, I., Claeys, H., De Rycke, R., Andriankaja, M., Van Aken, O., Van Breusegem, F., Fernie, A. and Inze, D., *Developmental stage specificity and the role of mitochondrial metabolism in the response of Arabidopsis leaves to prolonged mild osmotic stress.*, Plant Physiol **152(1)**, 226-244 (2010). 155

[289] Chan, Z., Grumet, R. and Loescher, W., *Global gene expression analysis of transgenic, mannitol-producing, and salt-tolerant Arabidopsis thaliana indicates widespread changes in abiotic and biotic stress-related genes.*, J Exp Bot **62(14)**, 4787-4803 (2011). 155, 157

[290] Pieterse, C., Leon-Reyes, A., Van Der Ent, S. and Van Wees, S.C., *Networking by small-molecule hormones in plant immunity.*, Nat Chem Biol **5(5)**, 308-316 (2009). 155

[291] Kim, T., *Plant stress surveillance monitored by ABA and disease signaling interactions.*, Mol Cells **33(1)**, 1-7 (2012). 155

[292] Dubois, M., Skirycz, A., Claeys, H., Maleux, K., Dhondt, S., De Bodt, S., Vanden Bossche, R., De Milde, L., Yoshizumi, T., Matsui, M. and Inze, D., *ETHYLENE RESPONSE FACTOR6 acts as a central regulator of leaf growth under water-limiting conditions in arabidopsis.*, Plant Physiol **162(1)**, 319-332 (2013). 155

[293] Tsuda, K., Sato, M., Stoddard, T., Glazebrook, J. and Katagiri, F., *Network properties of robust immunity in plants.*, PLoS Genet **5(12)**, e1000772 (2009). 155

[294] Thomas, J., Sepahi, M., Arendall, B. and Bohnert, H., *Enhancement of seed germination in high salinity by engineering mannitol expression in Arabidopsis thaliana*, Plant, Cell & Environment **18(7)**, 801-806 (1995). 157

[295] Karakas, B., Ozias-Akins, P., Stushnoff, C., Suefferheld, M. and Rieger, M., *Salinity and drought tolerance of mannitol-accumulating transgenic tobacco*, Plant, Cell & Environment **20(5)**, 609-616 (1997). 157

[296] Tarczynski, M., Jensen, R. and Bohnert, H., *Stress protection of transgenic tobacco by production of the osmolyte mannitol.*, Science **259(5094)**, 508-510 (1993). 157

[297] Zhifang G, L.W.H., *Expression of a celery mannose 6-phosphate reductase in* Arabidopsis thaliana *enhances salt tolerance and induces biosynthesis of both mannitol and a glucosyl-mannitol dimer*, Plant, Cell & Environment **26(2)**, 275-283 (2003). 157, 165

[298] Shen, B., Jensen, R. and Bohnert, H., *Increased resistance to oxidative stress in transgenic plants by targeting mannitol biosynthesis to chloroplasts.*, Plant Physiol **113(4)**, 1177-1183 (1997). 157

[299] Prabhavathi, V., Yadav, J., Kumar, P. and Rajam, M., *Abiotic stress tolerance in transgenic eggplant (solanum melongena l.) by introduction of bacterial mannitol phosphodehydrogenase gene*, Molecular Breeding **9(2)**, 137-147 (2002). 157

[300] Abebe, T., Guenzi, A., Martin, B. and Cushman, J., *Tolerance of mannitol-accumulating transgenic wheat to water stress and salinity.*, Plant Physiol **131(4)**, 1748-1755 (2003). 157

[301] Lewis, D.H. and Smith, D.C., *Sugar alcohols (polyols) in fungi and green plants: distribution, physiology and metabolism*, New Phytologist **66**, 143-184 (1967). 157

[302] Cheng, F., Zamski, E., Guo, W., Pharr, D. and Williamson, J., *Salicylic acid stimulates secretion of the normally symplastic enzyme mannitol dehydrogenase: A possible defense against mannitol-secreting fungal pathogens.*, Planta **230(6)**, 1093-1103 (2009). 157

[303] Williamson, J., Stoop, J., Massel, M., Conkling, M. and Pharr, D., *Sequence analysis of a mannitol dehydrogenase cdna from plants reveals a function for the pathogenesis-related protein ELI3.*, Proc Natl Acad Sci U S A **92(16)**, 7148-7152 (1995). 157

[304] Zamski, E., Guo, W., Yamamoto, Y., Pharr, D. and Williamson, J., *Analysis of celery (apium graveolens) mannitol dehydrogenase (MTD) promoter regulation in Arabidopsis suggests roles for MTD in key environmental and metabolic responses.*, Plant Mol Biol **47(5)**, 621-631 (2001). 157

[305] Jennings, D., Ehrenshaft, M., Pharr, D. and Williamson, J., *Roles for mannitol and mannitol dehydrogenase in active oxygen-mediated plant defense.*, Proc Natl Acad Sci U S A **95(25)**, 15129-15133 (1998). 158

[306] Jennings, D., Daub, M., Pharr, D. and Williamson, J., *Constitutive expression of a celery mannitol dehydrogenase in tobacco enhances resistance to the mannitol-secreting fungal pathogen Alternaria alternata.*, Plant J **32(1)**, 41-49 (2002). 158

[307] Voegele, R., Hahn, M., Lohaus, G., Link, T., Heiser, I. and Mendgen, K., *Possible roles for mannitol and mannitol dehydrogenase in the biotrophic plant pathogen Uromyces fabae.*, Plant Physiol **137(1)**, 190-198 (2005). 158

[308] Velez, H., Glassbrook, N. and Daub, M., *Mannitol biosynthesis is required for plant pathogenicity by Alternaria alternata.*, FEMS Microbiol Lett **285(1)**, 122-129 (2008). 158

[309] Govrin, E. and Levine, A., *The hypersensitive response facilitates plant infection by the necrotrophic pathogen Botrytis cinerea.*, Curr Biol **10(13)**, 751-757 (2000). 158

[310] Jobic, C., Boisson, A., Gout, E., Rascle, C., Fevre, M., Cotton, P. and Bligny, R., *Metabolic processes and carbon nutrient exchanges between host and pathogen sustain the disease development during sunflower infection by Sclerotinia sclerotiorum.*, Planta **226(1)**, 251-265 (2007). 158

[311] Solomon, P., Waters, O. and Oliver, R., *Decoding the mannitol enigma in filamentous fungi.*, Trends Microbiol **15(6)**, 257-262 (2007). 158

[312] Dulermo, T., Rascle, C., Chinnici, G., Gout, E., Bligny, R. and Cotton, P., *Dynamic carbon transfer during pathogenesis of sunflower by the necrotrophic fungus Botrytis cinerea: From plant hexoses to mannitol.*, New Phytol **183(4)**, 1149-1162 (2009). 158

[313] Dickman, M. and Mitra, A., *Arabidopsis thaliana as a model for studying Sclerotinia sclerotiorum pathogenesis.*, Physiol. Mol. Plant Pathol. **41**, 255-263 (1992). 158

[314] Wang, S. and Le Tourneau, D., *Mannitol biosynthesis in Sclerotinia sclerotiorum.*, Arch Mikrobiol **81(1)**, 91-99 (1972). 158

[315] Perchepied, L., Balague, C., Riou, C., Claudel-Renard, C., Riviere, N., Grezes-Besset, B. and Roby, D., *Nitric oxide participates in the complex interplay of defense-related signaling pathways controlling disease resistance to Sclerotinia sclerotiorum in Arabidopsis thaliana.*, Mol Plant Microbe Interact **23(7)**, 846-860 (2010). 158, 159

[316] Shiu, S. and Bleecker, A., *Receptor-like kinases from Arabidopsis form a monophyletic gene family related to animal receptor kinases.*, Proc Natl Acad Sci U S A **98(19)**, 10763-10768 (2001). 160, 168

[317] Shiu, S. and Bleecker, A., *Plant receptor-like kinase gene family: Diversity, function, and signaling.*, Sci STKE **2001(113)**, RE22 (2001). 160

[318] Albert, M., Jehle, A., Lipschis, M., Mueller, K., Zeng, Y. and Felix, G., *Regulation of cell behaviour by plant receptor kinases: Pattern recognition receptors as prototypical models.*, Eur J Cell Biol **89(2-3)**, 200-207 (2010). 160

[319] Lehti-Shiu, M.D., Zou, C., Hanada, K. and Shiu, S., *Evolutionary history and stress regulation of plant receptor-like kinase/pelle genes.*, Plant Physiol **150(1)**, 12-26 (2009). 160, 170

[320] Gish, L. and Clark, S., *The RLK/pelle family of kinases.*, Plant J **66(1)**, 117-127 (2011). 160

[321] Felix, G., Duran, J., Volko, S. and Boller, T., *Plants have a sensitive perception system for the most conserved domain of bacterial flagellin.*, Plant J **18(3)**, 265-276 (1999). 161

[322] Gomez-Gomez, L., Felix, G. and Boller, T., *A single locus determines sensitivity to bacterial flagellin in Arabidopsis thaliana.*, Plant J **18(3)**, 277-284 (1999). 161

[323] Gomez-Gomez, L. and Boller, T., *FLS2: An LRR receptor-like kinase involved in the perception of the bacterial elicitor flagellin in Arabidopsis.*, Mol Cell **5(6)**, 1003-1011 (2000). 161

[324] Dunning, F., Sun, W., Jansen, K., Helft, L. and Bent, A., *Identification and mutational analysis of Arabidopsis FLS2 leucine-rich repeat domain residues that contribute to flagellin perception.*, Plant Cell **19(10)**, 3297-3313 (2007). 161

[325] Schulze, B., Mentzel, T., Jehle, A., Mueller, K., Beeler, S., Boller, T., Felix, G. and Chinchilla, D., *Rapid heteromerization and phosphorylation of ligand-activated plant transmembrane receptors and their associated kinase BAK1.*, J Biol Chem **285(13)**, 9444-9451 (2010). 161

[326] Lu, D., Wu, S., Gao, X., Zhang, Y., Shan, L. and He, P., *A receptor-like cytoplasmic kinase, BIK1, associates with a flagellin receptor complex to initiate plant innate immunity.*, Proc Natl Acad Sci U S A **107(1)**, 496-501 (2010). 161

[327] Reyes, F., Buono, R. and Otegui, M., *Plant endosomal trafficking pathways.*, Curr Opin Plant Biol **14(6)**, 666-673 (2011). 161

[328] Robatzek, S., Chinchilla, D. and Boller, T., *Ligand-induced endocytosis of the pattern recognition receptor FLS2 in Arabidopsis.*, Genes Dev **20(5)**, 537-542 (2006). 161

[329] Lu, D., Lin, W., Gao, X., Wu, S., Cheng, C., Avila, J., Heese, A., Devarenne, T., He, P. and Shan, L., *Direct ubiquitination of pattern recognition receptor FLS2 attenuates plant innate immunity.*, Science **332(6036)**, 1439-1442 (2011). 161

[330] Irani, N. and Russinova, E., *Receptor endocytosis and signaling in plants.*, Curr Opin Plant Biol **12(6)**, 653-659 (2009). 161

[331] Falconer, R. and Collins, B., *Survey of the year 2009: Applications of isothermal titration calorimetry.*, J Mol Recognit **24(1)**, 1-16 (2011). 165

[332] Safina, G., *Application of surface plasmon resonance for the detection of carbohydrates, glycoconjugates, and measurement of the carbohydrate-specific interactions: A comparison with conventional analytical techniques. A critical review.*, Anal Chim Acta **712**, 9-29 (2012). 165

[333] Klepek, Y., Geiger, D., Stadler, R., Klebl, F., Landouar-Arsivaud, L., Lemoine, R., Hedrich, R. and Sauer, N., *Arabidopsis polyol transporter5, a new member of the monosaccharide transporter-like superfamily, mediates H+-symport of numerous substrates, including myo-inositol, glycerol, and ribose.*, Plant Cell **17(1)**, 204-218 (2005). 165

[334] Reinders, A., Panshyshyn, J. and Ward, J., *Analysis of transport activity of Arabidopsis sugar alcohol permease homolog AtPLT5.*, J Biol Chem **280(2)**, 1594-1602 (2005). 165

[335] Tordai, H., Banyai, L. and Patthy, L., *The pan module: The N-terminal domains of plasminogen and hepatocyte growth factor are homologous with the apple domains of the prekallikrein family and with a novel domain found in numerous nematode proteins.*, FEBS Lett **461(1-2)**, 63-67 (1999). 168

[336] Trotochaud, A., Hao, T., Wu, G., Yang, Z. and Clark, S., *The CLAVATA1 receptor-like kinase requires CLAVATA3 for its assembly into a signaling complex that includes kapp and a rho-related protein.*, Plant Cell **11(3)**, 393-406 (1999). 168

[337] Trotochaud, A., Jeong, S. and Clark, S., *CLAVATA3, a multimeric ligand for the CLAVATA1 receptor-kinase.*, Science **289(5479)**, 613-617 (2000). 168

[338] Heese, A., Hann, D., Gimenez-Ibanez, S., Jones, A., He, K., Li, J., Schroeder, J., Peck, S. and Rathjen, J., *The receptor-like kinase SERK3/BAK1 is a central regulator of innate immunity in plants.*, Proc Natl Acad Sci U S A **104(29)**, 12217-12222 (2007). 168

[339] Chinchilla, D., Zipfel, C., Robatzek, S., Kemmerling, B., Nurnberger, T., Jones, J., Felix, G. and Boller, T., *A flagellin-induced complex of the receptor FLS2 and BAK1 initiates plant defence.*, Nature **448(7152)**, 497-500 (2007). 168

[340] Takayama, S., Shimosato, H., Shiba, H., Funato, M., Che, F., Watanabe, M., Iwano, M. and Isogai, A., *Direct ligand-receptor complex interaction controls brassica self-incompatibility.*, Nature **413(6855)**, 534-538 (2001). 168

[341] Li, L., Foster, C., Gan, Q., Nettleton, D., James, M., Myers, A. and Wurtele, E., *Identification of the novel protein as a component of the starch metabolic network in Arabidopsis leaves.*, Plant J **58(3)**, 485-498 (2009). 173, 189, 192, 203

[342] Graf, A. and Smith, A., *Starch and the clock: The dark side of plant productivity.*, Trends Plant Sci **16(3)**, 169-175 (2011). 189, 193

[343] Stitt, M. and Zeeman, S., *Starch turnover: Pathways, regulation and role in growth.*, Curr Opin Plant Biol **15(3)**, 282-292 (2012). 189

[344] Seo, P., Kim, M., Ryu, J., Jeong, E. and Park, C., *Two splice variants of the IDD14 transcription factor competitively form nonfunctional heterodimers which may regulate starch metabolism.*, Nat Commun **2**, 303 (2011). 189, 192, 193, 203

[345] Espinoza, C., Degenkolbe, T., Caldana, C., Zuther, E., Leisse, A., Willmitzer, L., Hincha, D. and Hannah, M., *Interaction with diurnal and circadian regulation results in dynamic metabolic and transcriptional changes during cold acclimation in Arabidopsis.*, PLoS One **5(11)**, e14101 (2010). 193

[346] El-Lithy, M.E., Reymond, M., Stich, B., Koornneef, M. and Vreugdenhil, D., *The relation between plant growth, carbohydrates and flowering time in the Arabidopsis Landsberg erecta x Kondara recombinant inbred line population.*, Plant Cell Environ **33(8)**, 1369-1382 (2010). 193

[347] Pantin, F., Simonneau, T., Rolland, G., Dauzat, M. and Muller, B., *Control of leaf expansion: A developmental switch from metabolics to hydraulics.*, Plant Physiol **156(2)**, 803-815 (2011). 193

[348] Munns, R. and Tester, M., *Mechanisms of salinity tolerance.*, Annu Rev Plant Biol **59**, 651-681 (2008). 204

Résumé

Décoder la complexité de la variabilité naturelle pour la croissance et la réponse à l'environnement chez *Arabidopsis thaliana*.

Des génotypes adaptés à des environnements contrastés ont de grandes chances de se comporter différemment lorsqu'ils sont placés dans des conditions similaires et contrôlées, notamment si leur sensibilité aux signaux environnementaux et/ou leur croissance intrinsèque sont limitées à différents niveaux. De ce fait, la variabilité observée dans les populations naturelles peut être utilisée comme une source illimitée de nouveaux allèles ou gènes pour l'étude des bases génétiques de la variation des traits quantitatifs. Mon travail de doctorat a consisté en l'analyse de la variabilité naturelle pour la croissance et la réponse è l'environnement chez Arabidopsis thaliana. Le but des approches de génétique quantitative est de comprendre comment la diversité génétique et épigénétique contrôle la variabilité phénotypique observée dans les populations à différentes échelles, au cours du développement et sous différentes contraintes environnementales. De plus, ces analyses ont pour objectif de comprendre comment les processus adaptatifs et démographiques influencent la fréquence de ces variants dans les populations en fonction de leur environnement local. Ainsi, l'étude de la variabilité naturelle peut être appréhendée en utilisant diverses approches, de la génétique et des méthodes de biologie moléculaire aux études écologiques et évolutives. Au cours de mon doctorat, j'ai eu la chance de travailler sur plusieurs de ces aspects au travers de trois projets indépendants qui exploitent tous la variabilité naturelle d'*A. thaliana*.

Le premier projet a consisté en l'analyse du pattern de polymorphisme observé dans des populations d'A. thaliana au gène MOT1 qui code pour un transporteur de molybdate (la forme assimilable du molybdène (Mo), un micro-élément essentiel) et qui est responsable d'une partie des variations de croissance et de fitness observées à l'échelle de l'espèce en fonction de la disponibilité en Mo des sols. J'ai montré à différentes échelles géographiques que le pattern de polymorphisme à MOT1 ne reflète pas une évolution neutre mais présente plutôt des traces de sélection diversifiante. Ce travail a contribué à renforcer l'hypothèse selon laquelle des mutations au niveau du gène MOT1 pourraient avoir été sélectionnées dans certaines populations pour faire face aux niveaux élevés de Mo observés dans certains sols et potentiellement délétères malgré leur effet négatif sur des milieux pauvres en Mo.

Le deuxième projet portait sur la caractérisation et l'analyse fonctionnelle de deux récepteur-kinase putatifs (RLK) identifiés de part leurs effets sur la croissance foliaire spécifiquement en réponse à un stress induit par du mannitol mais pas sous d'autres contraintes osmotiques. La fonction de ces récepteurs chez A. thaliana -qui n'est pas connu pour produire du mannitol- peut paraître intrigante. Les différentes expériences réalisées au cours de cette thèse nous ont cependant permis de construire un modèle selon lequel ces récepteurs pourraient être activés par le mannitol produit par certains pathogènes tel que les champignons et participer aux réponses de défense de la plante.

Le troisième projet a été réalisé en collaboration avec l'équipe de Michel Vincentz (CBMEG, Brésil) et de Vincent Colot (IBENS, Paris) et consiste en l'analyse de l'occurrence de variants épigénétiques naturels au gène QQS dans différentes populations d'Asie Centrale et de leurs possibles conséquences phénotypique et adaptative.

En conclusion, l'analyse des variants génétiques et épigénétiques naturels à l'origine des variations de biomasse en interaction avec l'environnement permet de comprendre comment l'évolution façonne la variabilité naturelle.

Mots clefs : Variabilité naturelle, *Arabidopsis thaliana*, accessions, RILs, croissance foliaire, environnement, stress abiotique et biotique, génétique quantitative, quantitative trait locus (QTL).

Abstract

Decoding the complexity of natural variation for shoot growth and response to the environment in *Arabidopsis thaliana*

Genotypes adapted to contrasting environments are expected to behave differently when placed in common controlled conditions, if their sensitivity to environmental cues or intrinsic growth behaviour are set to different thresholds, or are limited at distinct levels. This allows natural variation to be exploited as an unlimited source of new alleles or genes for the study of the genetic basis of quantitative trait variation. My doctoral work focuses on analysing natural variation for shoot growth and response to the environment in *A. thaliana*. Natural variation analyses aim at understanding how molecular genetic or epigenetic diversity controls phenotypic variation at different scales and times of plant development and under different environmental conditions, and how selection or demographic processes influence the frequency of those molecular variants in populations for them to get adapted to their local environment. As such, the analysis of *A. thaliana* natural variation can be addressed using a variety of approaches, from genetics and molecular methods to ecology and evolutionary questions. During my PhD, I got the chance to tackle several of those aspects through my contributions to three independent projects which have in common to exploit *A. thaliana* natural variation.

The first one is the analysis of the pattern of polymorphism from a set of 102 *A. thaliana* accessions at the *MOT1* gene coding for a molybdate transporter (an essential micronutrient) and responsible for contrasted growth and fitness among accessions in response to Mo availability in the soil. I showed at different geographical scales that *MOT1* patt@ern of polymorphisms is not consistent with neutral evolution and shows signs of diversifying selection. This work helped reinforce the hypothesis that in some populations, mutations in *MOT1* have been selected to face soils rich in Mo and potentially deleterious despite their negative effect on Mo-limiting soils.

The second project consists in the characterisation and functional analysis of two putative receptor-like kinases (RLKs) identified from their effect on shoot growth specifically under mannitol-supplemented media and not in response to other osmotic constraints. The function of such RLKs in *A. thaliana*, which is not known to synthesize mannitol was intriguing at first but, through different experiments, we built the hypothesis that those RLKs could be activated by the mannitol produced by some pathogens such as fungi and participate to plant defensive response.

The third project, in collaboration with Michel Vincentz's team from CBMEG (Brasil) and Vincent Colot (IBENS, Paris), consists in the analysis of the occurrence of natural epigenetic variants of the *QQS* gene in different populations from Central Asia and their possible phenotypic and adaptive consequences.

Overall, these analyses of the genetic and epigenetic molecular variation leading to the biomass phenotype(s) in interaction with the environment provide clues as to how and where in the pathways adaptation is shaping natural variation.

Keywords: Natural variation, Arabidopsis thaliana, accessions, RILs, shoot growth, environment, abiotic and biotic stress, quantitative genetics, quantitative trait locus (QTL).