Première exploration du paysage génomique intraspécifique de l'adaptation chez les plantes non vasculaires : le cas de l'hépatique Marchantia polymorpha

Chloé Beaulieu

Doctorat de l'Université de Toulouse

préparé à l'Université Toulouse III - Paul Sabatier

Première exploration du paysage génomique intraspécifique de l'adaptation chez les plantes non vasculaires : le cas de l'hépatique Marchantia polymorpha

---

Thèse présentée et soutenue, le 1 mars 2024 par

Chloé BEAULIEU

École doctorale
SEVAB - Sciences Ecologiques, Vétérinaires, Agronomiques et Bioingenieries

Spécialité
Developpement des plantes, interactions biotiques et abiotiques

Unité de recherche
LRSV - Laboratoire de Recherche en Sciences Végétales

Thèse dirigée par
Maxime BONHOMME et Pierre-Marc DELAUX

Composition du jury
M. Stig ANDERSEN, Rapporteur, Aarhus University
Mme Isabel MONTE GRONDONA, Rapporteure, Universität Tübingen
M. Bruno CONTRERAS MOREIRA, Rapporteur, Consejo Superior de Investigaciones Cientificas
M. Jean-Philippe GALAUD, Examinateur, Université Toulouse III - Paul Sabatier
M. Maxime BONHOMME, Directeur de thèse, Université Toulouse III - Paul Sabatier
M. Pierre-Marc DELAUX, Co-directeur de thèse, CNRS Occitanie-Ouest

Membres invités
M. Cyril Libourel, Ecole d'ingénieurs de Purpan

Abstract

Plant adaptation to a terrestrial life 500 million years ago played a major role in the evolution of life on Earth. Plants still play a crucial role at present time, as bases of most ecosystems, and consequently base of all human civilizations. Understanding their adaptation to past and present modification of their living conditions is a key to understand the past and be able to respond to future agricultural challenges. Plants sciences made great advances in understanding plant response to their environment, but most studies focused on the angiosperm lineage which contains crops. Nevertheless, to get a broader picture of land plant (Embryophytes) adaptation to various conditions, in the framework of 500 million years of evolution on land, it is essential to study other land plant lineages. In line with this logic, this work will focus on the non-vascular plant model *Marchantia polymorpha*, whose lineage diverged from vascular plants around 480 million years ago. We developed an intraspecific diversity dataset that allowed us to uncover some mechanisms of adaptation in *M. polymorpha*. Analyses of selection signatures on genes enabled us to distinguish conserved functions under strong purifying selection from variable ones undergoing balancing selection or selective sweeps. Using this intraspecific diversity dataset, genome-wide associations studies (GWAS) could be performed on the response of *M. polymorpha* to climatic conditions but also to the fungal pathogen *Colletotrichum nymphaeae* (biotic stress). Finally, a gene-based pangenome was built and allowed identifying genes with a presence-absence variation between accessions, that are often associated with stress response and local adaptation. Crossing these three approaches, we found gene families that seem involved in *M. polymorpha* response to stresses. Among them can be cited the terpene synthases, the peroxidases, the NBS-LRR (NLR), the lectins, the lipoxygenases or the polyphenol oxidases. Most of these functions are shared with other land plants, showing that most general mechanism of adaptation are quite conserved in Embryophytes. Nevertheless, most of these gene families displayed lineage-specific characteristics, such as specific genes, family expansions or horizontal gene transfer, that differentiated gene family organization in Marchantia from the one known in angiosperms. Taken together, these results show that land plants share most of their mechanisms of adaptation, inherited from their last common ancestor, and that these general functions underwent lineage-specific modifications, that can hint at the different constraints that shaped the different lineages of land plants.

Keywords: bryophyte, adaptation, pangenome , phylogenomics, GWAS, land plants

Resumé

L'adaptation des plantes à la vie terrestre il y a 500 millions d'années a joué un rôle majeur dans l'évolution de la vie sur Terre. Les plantes jouent toujours un rôle crucial à l'heure actuelle, en tant que base de la plupart des écosystèmes et, par conséquent, base de toutes les civilisations humaines. Comprendre leur adaptation aux modifications passées et présentes de leurs conditions de vie est une clef pour comprendre le passé et être capable de répondre aux défis agricoles futurs. Les sciences végétales ont fait de grands progrès dans la compréhension de la réponse des plantes à leur environnement, mais la plupart des études se sont concentrées sur la lignée des angiospermes, qui comprend les plantes cultivées. Néanmoins, pour avoir une vision plus large de l'adaptation des plantes terrestres (Embryophytes) à diverses conditions, dans le cadre de 500 millions d'années d'évolution sur la terre ferme, il est essentiel d'étudier d'autres lignées de plantes terrestres. Dans cette logique, ce travail se focalisera sur la plante modèle non vasculaire *Marchantia polymorpha*, dont la lignée a divergé des plantes vasculaires il y a environ 480 millions d'années. Nous avons développé un ensemble de données sur la diversité intraspécifique qui nous a permis de mettre en évidence certains mécanismes d'adaptation chez *M. polymorpha*. L'analyse des signatures de sélection sur les gènes nous a permis de distinguer les fonctions conservées soumises à une forte sélection purificatrice des fonctions variables soumises à une sélection équilibrante ou à des balayages sélectifs. En utilisant cet ensemble de données sur la diversité intraspécifique, des études d'association à l'échelle du génome (GWAS) ont pu être réalisées sur la réponse de *M. polymorpha* aux conditions climatiques mais aussi au champignon pathogène *Colletotrichum nymphaeae* (stress biotique). Enfin, un pangénome basé sur les gènes a été construit et a permis d'identifier des gènes avec une variation de présence-absence entre les accessions, qui sont souvent associés à la réponse au stress et à l'adaptation locale. En croisant ces trois approches, nous avons trouvé des familles de gènes qui semblent impliquées dans la réponse de *M. polymorpha* aux stress. Parmi elles, on peut citer les terpènes synthases, les peroxydases, les NBS-LRR (NLR), les lectines, les lipoxygénases ou les polyphénols oxydases. La plupart de ces fonctions sont partagées avec d'autres plantes terrestres, ce qui montre que la plupart des mécanismes généraux d'adaptation sont assez conservés chez les Embryophytes. Néanmoins, la plupart de ces familles de gènes présentent des caractéristiques propres à la lignée, telles que des gènes spécifiques, des expansions de famille ou des transferts horizontaux de gènes, qui différencient l'organisation des familles de gènes chez les Marchantia de celle connue chez les angiospermes. L'ensemble de ces résultats montre que les plantes terrestres partagent la plupart de leurs mécanismes d'adaptation, hérités de leur dernier ancêtre commun, et que ces fonctions générales ont subi des modifications spécifiques à chaque lignée, dépendant des différentes contraintes qui ont façonné les lignées de plantes terrestres.

Keywords: bryophyte, adaptation, pangénome , phylogénomique, GWAS, plantes terrestres

Remerciements (Acknowledgments)

Tout d'abord, je tenais à remercier mes deux directeurs de thèse pour m'avoir permis d'effectuer cette thèse surprise comme j'aime à l'appeler. J'ai vraiment apprécié les trois (et quelques…) dernières années que j'ai passées sur ce projet, encadrée par vous ! Vous formez vraiment un super duo d'encadrant, aussi bien par la complémentarité de vos domaines d'expertise (scientifiques mais aussi humains !) que par l'excellente communication que vous entretenez entre vous (de ce que j'ai pu comprendre, ne pas avoir à faire le pigeon voyageur est un privilège que peu de doctorants co-encadrés ont). Au-delà de votre duo de choc, vous m'avez chacun beaucoup apporté.

Merci Maxime pour ton suivi quasi-quotidien, ton écoute et ta bonne humeur ! Tu es vraiment le manager parfait entre dédramatisation et organisation ! Tu as même réussi à rendre fun et enthousiasmantes des choses qui auraient pu me faire paniquer (comme faire une thèse par exemple…), juste en les présentant bien. Merci pour tous les feedbacks, toujours tournés positivement, qui ont toujours réussi à me relancer, même quand je pouvais en avoir marre. Merci aussi de m'avoir toujours fait me sentir ton égale, tu as un réel talent pour manager sans jamais commander ! Merci, aussi, pour toutes les discussions scientifiques, qui étaient de vrais échanges, c'était un plaisir à chaque fois ! Je pense que peu de gens ont la chance d'avoir un directeur de thèse aussi impliqué et aussi bienveillant que toi.

Merci PM pour ton enthousiasme à toute épreuve (et je pèse mes mots) et tes connaissances si vastes. Tu arrives toujours à t'investir dans le projet malgré ton agenda de ministre, et à nous faire profiter de tes multiples talents : community manager/influenceur, élagueur de papier, bibliothèque vivante. Comme Max, tu es un encadrant avec lequel on se sent à l'aise et à qui j'ai pu parler de mes inquiétudes de tous types, liées à ce projet (hélas pour toi, sûrement, qui a dû confronter ton optimisme débordant à mon pessimisme à peine contenu) ! Merci aussi pour la bonne ambiance que tu insuffles à l'équipe et pour les activités que tu essaies de mettre en place, ces trois années en EVO ont été un plaisir en partie grâce à cela !

Merci au programme 80 prime du CNRS qui a permis de financer ce projet et m'a donc globalement permis d'étudier Marchantia tout en ayant un toit sur la tête !

Merci à la team bioinfo, le sang de la veine de ma thèse (et aussi de la vie), sans vous rien n'aurait été possible (dans un sens totalement littéral). Merci à Cyril de m'avoir aidé à porter ce projet, toutes nos discussions ont été tellement intéressantes que ce soit sur Marchantia ou pour des débriefs de vie. Concrètement, sans toi, je serais roulée en boule dans un coin de mon bureau à douter de toutes les techniques que j'emploie, et de tout ce que j'ai fait en général. Tu es un conseiller pragmatique et déculpabilisant qui m'a sauvé de mille crises existentielles ! Merci aussi d'avoir fait de moi une pro de R, n'en déplaise à Jean. D'ailleurs merci à toi Jean pour tes conseils toujours avisés et très précis, tu m'as mis dans les rails du travail bien fait (même si je ne serais clairement jamais à ton niveau). Merci pour cette énergie folle et ta gentillesse à toute épreuve (il n'y a que quelqu'un d'aussi hyperactif que toi qui puisse être aussi investi et serviable avec tout le monde !). Merci à Campu, ma binôme, ma demi sœur de gay dads, l'algue à mon fungi. Tu remarqueras bien que je fais de mon mieux pour ne pas t'abandonner trop tôt ! Merci pour tous les pro tips info quand j'étais perdue, les débriefs de nos week-ends, mais aussi et surtout pour le soutien indéfectible au quotidien (je ne pourrais

jamais assez vous remercier avec Cyril, pour les petites séances de psy du début 2023). C'est une belle amitié qui a commencé avec ton stage de M2, mastercroute ! Merci aussi à Paoline, notre bioinfo invitée, pour avoir insufflé un vent nouveau dans notre équipe (j'ai adoré les lamantins plots même s'ils n'ont pas été des plus concluants), pour tes histoires légendaires, tes milles activités et pour ta collection de sculptures d'animaux (sans toi je n'aurais jamais fait de poterie et ça aurait été une véritable tragédie). J'ai hâte de te revoir jouer à des jeux de société de niche et admirer ta déco, mais en attendant régale toi dans ta nouvelle vie Yale-ienne, tu le mérites tellement ! A big thank also to the latest bioinfo addition, Fabiano(x), for your incredible cuteness and ability to produce weird noises. It has been so fun to have you in the bioinfo office and being small lost animals together (bioinfo newbies represent! loving the existential crisis conversations)! Merci à Inès, tu as été une parfaite addition au bureau bioinfo bien que tu appartiennes à l'autre monde du labo ! Ta passion de la bouffe, tes phrases motivationnelles et tes périples chaque week-end m'ont régalée, merci pour toute cette énergie et cette bienveillance ! Merci aussi à Madeleine la nouvelle venue du bureau bioinfo pour tes folles histoires et ta gentillesse !

Merci à l'équipe EVO pour ces 3 années de travail dans la bonne humeur, notamment aux visiteurs réguliers du bureau de la bouffe comme Tatiana et Mélanie (bon de base, c'était pour Tic et Tac, mais vous avez continué de venir de temps à autre après leurs départs), CamilleG (team cam-algues rpz !), Christophe (le nouveau voisin qui squatte la cafetière), mais aussi le reste de la team : Soizic (merci pour les crises existentielles sur l'écologie pendant les lab meetings, c'était clairement mes prefs !), Ayla (la sagesse enveloppée dans 20 pulls en hiver), Eve brune et Eve blonde (désolé mais oui je vous différencie comme ça dans ma tête), Lucie, Laurena, Tifenn, Léa, Leonardo, Baptiste C, Katharina, Nico, Dominique, Corinne, Viriginie…, et merci à Karima pour ton travail sur la GWAS Marchantia et pour ton investissement dans la vie du labo !

Merci Hervé pour ta collaboration sur le projet depuis ton petit coin d'Ariège et ton regard complémentaire sur les aspects phylogénétiques chez Marchantia !

Merci Hélène pour tout ton travail sur la V1 du pangénome et pour tous les petits conseils bioinfos que tu as pu me donner, ainsi que pour l'énorme travail de maintenance informatique que tu fais pour le LRSV !

Merci aux membres de mon comité de thèse, Nathalie Chantret et Fabrice Roux pour ces trois ans de suivi, vos conseils et votre bienveillance. Un merci additionnel à Fabrice pour ton aide sur le traitement des données phénotypiques que la GWAS.

Merci à Fabien Mounet de m'avoir permis d'effectuer des enseignements, et surtout merci de m'avoir fait confiance pour le TP ACP, c'était clairement l'expérience d'enseignement la plus plaisante et enrichissante que j'ai eu l'occasion d'avoir !

Merci à l'équipe de la Bioinfo genotoul qui sont incroyablement efficaces et m'ont sauvé à de multiples reprises après des échecs d'installation de logiciel sur mon ordi ! Sans le cluster, il n'y aurait peut-être eu que deux analyses dans ma thèse !

Merci au service administratif du LRSV, Catherine Deprey, Odile Barbier, David Andrieux et Sabine Leygues qui m'ont guidée dans d'innombrables démarches administratives !

Merci à Clemsouille pour ces trois ans au même rythme ou presque (oupsi tu m'as un peu semée sur la fin) et cette super complicité ! J'ai adoré toutes nos soirées à St Mich, les quelques escapades en concert, les rdv mioumiou/scrounchscrounch et compagnie et aussi les meetings couloirs impromptus qui durent parfois plus longtemps que prévu ! J'espère que tu vas te régaler dans ta nouvelle étape de vie et que tu repasseras à Toutou quand même de temps en temps !

Je voudrais aussi remercier tous les gens qui m'accompagnent depuis ces 26 dernières années d'existence, et sans le support de qui je n'aurais pas pu faire des choix d'apparence douteux, comme faire une thèse !

D'abord merci à la légendaire coloc pour cette vie de famille qu'on a eue au cours des dernières années. J'ai toujours pensé que je ne pourrais pas vivre en coloc, mais finalement, j'ai adoré vivre avec vous (même si je trouve toujours que vous ne mettez pas bien les couverts dans le lave-vaisselle), vous avez vraiment été ma base, surtout pendant les confinements, où on était les seules interactions sociales les uns des autres. On a créé tellement de souvenirs dans cette petite maison des 36 ponts, que ce soit dans la vie de tous les jours où dans les moments où elle devenait le QG officiel de la Flav ! Merci à Axelito pour toutes ces dernières années de twin life (ça commence à faire après 7 ans de vies parallèles). Je suis super contente de partager encore une expérience d'études avec toi, mais aussi les séries (oupsi Glee rest in peace), les debriefs de la rage, les livres, les crises existentielles. Tu m'as inspiré à me surpasser, mais j'aurais espéré t'apprendre à te sous-passer un peu plus… Quoi qu'il en soit, tu restes un grand scientifique polyvalent qui y comprend beaucoup plus à Marchantia que moi à PDL1 (oupsi)! Merci à Sofi pour cette énergie chaotique (viiiiiite) et solaire. Je parle beaucoup plus bizarrement qu'en arrivant à la coloc, mais je suis quand même tellement heureuse d'avoir partagé ces 3 dernières années avec toi. Tu as un don pour rassembler les gens, ce qui m'aura permis de ne pas devenir une ermite, et ta folie est la meilleure source de divertissement du monde ! Avant de te rencontrer, je n'avais jamais vu quelqu'un qui percevait la vie si différemment de moi, et maintenant je suis sûre que je ne pourrais plus m'en passer ! Merci SAS. Merci aussi à Dyko, it was so nice to get the chance to properly know you by living together and watching you having crazy projects each day. I have learned so much about 3D printing, bikes, the street market, pepper plants, NFT and many other things… and also about seing life in a chill way! Hope you'll make me discover your home state one day! Et merci aussi à Guigz, mon fidèle compère du bas, pour ta culture à toute épreuve (honnêtement choquée de ta connaissance des ères géologiques j'aurais dû t'embaucher pour écrire mon intro !!), tes talents culinaires (qui n'infirment pas les miens soyons bien clairs !), ta sensibilité et tes pics de nettoyage, rares mais efficaces ! Ça a été un plaisir de plus te découvrir à la coloc, et j'espère que tu vas continuer à explorer la vie pour trouver ton épanouissement !

Comment remercier la coloc sans remercier le reste de la Flav, qui sont tous en réalité des colocs honoraires et qui font mon bonheur chaque jour. Ça fait maintenant pas mal de temps que notre petit groupe se suit et évolue au gré des nouveaux arrivés, mais c'est toujours un plaisir absolu d'être avec vous ! Merci Fafa d'être la actu Toulouse la plus solaire, la peintre la

plus douée (tu m'apprendras) et une 5e coloc de choc. Je ne sais pas ce que l'avenir te réserve, mais je suis sûre que ce sera très apaisé et très artistique ! Merci Gathou d'être la douceur incarnée, tu es clairement dans mon top 5 des choses les plus mignonnes du monde (même si tu sais être impressionnante quand tu t'énerves) et j'ai toujours admiré ta décontraction (oui oui passer son temps à perdre des trucs aussi random que des bols dans sa voiture, c'est mon objectif de vie). Merci Ranou, ma première amie de l'INSA, pour la folie et la sagesse qui me régalent toujours autant (et aussi pour ta production de meilleures expressions du monde !). Tout le monde n'a pas la chance comme moi d'avoir dans sa vie une icône de mode qui sait faire la fractale ! Merci Cam pour les séances de danse endiablées, ton dynamisme à toute épreuve et ton caractère émotionnellement enceinte. J'ai adoré te découvrir et te voir évoluer au cours des dernières années et j'ai hâte de la suite avec plus d'aventures Royannaises, de debrief de nos lifes et de tisanes à la CAF ! Merci Paulo pour les imitations croquées, les tapes dans les mains et ta stabilité à toute épreuve (émotionnelle, aussi, hein, ce n'est pas qu'une question de grands pieds). Merci Flovino pour ton enthousiasme de jeune chiot, et ton implication infaillible ! Je suis tellement contente que tu fasses moult incursions dans le sud et qu'on ne te perde pas de vue malgré ton choix douteux d'aller vivre à Paris ! Merci Oranou pour les super discussions et les livres (que j'ai toujours un jour faudra que je te les rende quand même), et ton aura (idem que pour Paul, ce n'est pas qu'une histoire de cheveux même s'ils sont sacrément stylés !). Merci Nathan pour tes récits mouvementés, c'est toujours sportif quelle que soit l'activité, et tu nous fais voyager vers de nouveaux horizons ! Tu m'impressionnes de fou, mais essaies quand même de pas mourir stp ! Merci Auke pour tes moments de folie du rouge, ta solidarité au buffet (et dans les conséquences du buffet) et ton bon goût à toute épreuve. C'est un plaisir de te voir te livrer au fil des années et devenir un animal sauvage qui détruit les apparts ! Merci Max pour ta folie d'un autre type de rouge, et pour toutes les expériences que tu nous partages, c'est toujours un plaisir de réfléchir à la vie avec toi. Grâce à toi, je ne perds pas totalement contact avec mon côté babos ! Merci Hippo, le petit nouveau, mais pas des moindres ! Tu as été un super voisin, un incroyable showman (de violon, de stand up et de rédaction de thèse of course). Même si tu es à présent parti dans le pays de « la donna mangia una mela », tu seras toujours le bienvenu pour passer à la coloc avec ton sac de rando à moitié vide (ou pas, apparemment, ce sera una valigia con pesto je t'attends au tournant) ! Merci David, l'autre petit nouveau, d'être notre deuxième américain d'adoption ! It was a pleasure hanging with you this past 2 years, your sense of fashion, your humour and your kindness are on point, don't change anything! Merci Valou, le calme incarné dans un style parfait. Les déménagements de tes parents m'ont fait voyager plus que n'importe quelles vacances. C'est toujours un plaisir de te découvrir à la coloc en train de manger les risottos d'Axel ! Merci Nico pour les bons moments partagés, notamment pendant le road trip en 5A, mais aussi les soirées et petits moments ces dernières années. Tu es un véritable atout chill et pépite verbale ! Merci Nono pour les entraînements (même si j'ai rajquit tu es une super coach) et ta vision de la vie toujours intéressante à discuter ! Merci Rémy pour ton regard sur les choses, tu m'as montré qu'on pouvait faire une thèse sans avoir envie de crever de stress ! Merci Mathouprairie, notre Lyonnaise pref pour ton énergie de lapin enragé, qui te permet de mener les projets les plus fous et les soirées les plus folles aussi !

Je suis tellement contente d'appartenir à cette famille de coupains, toutes nos activités ont formé la plupart des meilleurs moments de ma vie Toulousaine, alors merci les ptits loups !

Merci aussi à Nélonie, acide aminé essentiel de la triade ! Même si on s'est fait avoir par les thèses et qu'on a des petits problèmes de gestion d'emploi du temps à chaque fois, c'est tellement un plaisir de te voir ! Tu me régales avec ton énergie chaotique et tes histoires incroyables. J'espère qu'on arrivera à un peu mieux gérer nos lifes à l'avenir pour pouvoir se refaire des petites sorties musées ou bouffe, ou tout simplement se poser tranquille à discuter de nos vies.

Merci à la team de Grégré, vous êtes mes sœurs d'autres mères (oui, oui, même toi Guigui) et je suis tellement contente de continuer à grandir avec vous ! Merci pour tout le soutien depuis tout ce temps, je ne sais pas comment j'aurais fait sans vous, vous êtes mes racines, ou plutôt mon ancre si on veut être dans le thème de la région. Merci pour toutes les discussions, les escapades, les soirées, les arrêts pipis intempestifs et les fous rires ! Merci à toi Meli d'être ma vrai-fausse jumelle depuis quasiment 20 ans ! C'est drôle parce que je pense qu'on se ressemble peut-être encore plus maintenant qu'à l'époque où tout le monde nous confondait h24 ! C'est toujours un plaisir de te voir que ce soit pour explorer de nouveaux endroits ou pour refaire le monde. Merci à Romy, tu es le yang à mon yin qui m'a fait découvrir que la communication sur les émotions, les chiens et le code, c'était globalement plutôt cool ! On a partagé tant de chose, de tes douches à des aventures Toulousaines et j'ai hâte que ça continue encore pour de nombreuses années (bien que ta nouvelle salle de bain m'ait l'air un peu moins confortable pour les débriefs). Merci à Mathou ma force de la nature préférée ! Ton énergie, tes talents organisationnels et ta bonne humeur n'ont d'égal que ta capacité à t'endormir absolument partout. Même avec nos fast lifes on a quand même trouvé le moyen de passer plein de moments privilégiés ces dernières années, avec des longues discussions bien trop tard et des aventures parfois un peu trop folles pour la trouillarde que je suis et j'ai adoré (sauf la trottinette électrique, ça plus jamais !). Hâte de revenir habiter chez toi à Marseille, et tu sais que tu es toujours la bienvenue quand tu passes à Toutou (surtout maintenant que tu sais où j'habite hein 😉). Merci Guigui, l'avocat du diable aux excellents goûts musicaux (je dis pas du tout ça parce que j'ai un peu les mêmes). Je suis tellement contente de voir que notre amitié de petits intellos qui se prêtent des livres continue encore aujourd'hui, à base de méga soirées debrief (tu es clairement notre consultant soirées filles :'), de balades dans la colline et de chalets à l'accès plus qu'accidenté (en vrai c'était trop fou !). J'espère que tu vas te régaler en Argentine, faire plein de belles rencontres et que tu auras plein de réflexions philosophiques sur la vie à nous partager !

Merci aussi à vos familles qui font aussi au final partie de ma vie (et qui me l'ont montré de la plus touchante des façons il n'y a pas tant de temps). Merci aux Regnier de toujours me laisser venir chez vous-même si je mange tout ce qui traîne et que je casse vos placards. Merci aux Girard pour tous les goûters impromptus. Merci aux Aimar pour l'accueil toujours hyper chaleureux.

Merci à Doud, le petit électron libre, de ne jamais être loin de mon champ gravitationnel, depuis nos 2 ans. A chaque fois qu'on se voit, c'est comme revenir à la maison ! Merci pour toutes les discussions super enrichissantes et tous ces moments passés dans toutes nos villes, Paris,

Toulouse, Bordeaux… et même dans la forêt de grégré ! Hâte de continuer à cheminer en parallèle, même si cette histoire d'océan entre la France et le Canada ne facilite pas les choses (edit : bon bah ça va en fait !).

Merci à Baptou pour toutes nos aventures itinérantes dans la nature, pour nos séances de cuisine en team (qui peuvent tourner au fight parce que la cuisine c'est sérieux !), pour toute cette culture que je n'aurais jamais cru avoir (#F1 et D&D) et surtout pour l'acceptation sans conditions ! Je suis tellement contente que nos chemins se soient recroisés. J'ai hâte de la suite avec encore plus de nourriture, de blagues plus ou moins drôles (mais souvent super drôle, sinon comment tu te ferais autant rire ?), de débats et d'aventures !

Merci aussi à la Sarrette family pour votre accueil, mes séjours Saint-Orennais sont toujours fort sympathiques !

Merci à mes parents de m'avoir élevée avec amour et de m'avoir accompagnée dans mes choix. Merci Maman pour ton écoute, ton soutien constant et tes talents de DIY que tu m'as transmis. Merci Papa pour ton calme, ton organisation et ta passion des voyages qu'on partage. J'ai une chance immense de vous avoir comme parents et aussi un peu comme amis ! Si je m'épanouis autant dans ma vie, c'est grâce à vous. Merci Quentin Blake d'être mon opposé, mais quand même vachement stylé ! Je suis super fière du jeune homme que tu es devenu (j'ai l'air d'une vieille quand je dis ça, j'en ai bien conscience). J'espère que tu vas continuer à t'épanouir et à avoir tes quarts d'heure de folie (bien que maintenant tu aies un peu trop de force pour que ça soit gérable !)

Merci à la famille des Poiriers (Bertrand et Gerard et mamie Haricot) et à tous ceux en Mayenne qui m'ont vu grandir année après année. On ne s'est pas vus très souvent, surtout ces dernières années, mais j'ai passé de tellement beaux moments avec vous !

Et enfin merci à toi mamie Colette, je suis convaincue que sans ton éducation, à la fois scolaire et sur la vie, rien de tout ça n'aurait été possible ! Tu es mon icône et tu m'auras impressionnée jusqu'au bout par ta force. C'est à toi que je dédie cette thèse.

List of abbreviations

ABA: Abscissic acid

AMS: Arbuscular Mycorrhizal Symbiosis

GEA: Genome-Environment Association

GWAS: Genome-Wide Association Study

HGT: Horizontal Gene Transfer

HOG: Hierarchical Orthogroup

JA: Jasmonic Acid

LD: Linkage Disequilibrium

LoF: Loss of Function

MTPSL: Microbial Terpene Synthase Like

MYA: Million Years Ago

NLR: Nucleotide binding domain Leucine rich Repeat proteins

HNL: αβ-Hydrolase NLR

CNL: Coil-coil NLR

TNL: TIR NLR

RNL: RPW8 NLR

PAV: Presence Absence Variation

PCA: Principal Component Analysis (PC: principal component)

PR: Pathogenesis Related protein

QTL : Quantitative Trait Locus

SA : Salicylic Acid

SNP: Single Nucleotide Polymorphism

SFS: Site Frequency Spectrum

SV: Structural Variation

TF: Transcription Factor

TPS: Terpenes Synthases

Table of content

Introduction

I) The importance of genetic diversity in evolution 2

II) The study of genetic diversity: history and tools 5

1) A short history of genetic diversity studies 5

2) SNP-based methods to study population evolution 8

a) Studies of demographic history 8

b) Studies of selection footprints 9

c) Examples of applications in plants 10

d) Quantitative genetics: crossing genetic polymorphisms with phenotypes (or other traits) 12

3) To complement the SNPs: Pangenomic studies 17

III) Land plant diversity and adaptations 22

1) The diversity of land plants 23

2) Common "tools" to thrive out of water 25

3) A diversity of clade-specific innovations associated with the diversification of land plants 28

IV) Perks of studying a model bryophyte: *Marchantia polymorpha* 30

1) The importance of studying bryophytes (and, more generally, other understudied plant lineages) 30

2) *Marchantia polymorpha* as a model bryophyte 32

V) Thesis objectives 38

Chapter I: *Marchantia polymorpha* diversity dataset and population genomics analyses

I) The *M. polymorpha* diversity dataset 42

1) Description of the collection 42

a) Collaborative sampling of a broad collection of accessions 42

b) Sequencing data on the accessions 45

2) Single Nucleotide Polymorphism (SNP) dataset construction 47

a) Mapping on the reference genome 47

b) SNP calling 48

c) Genome-wide patterns of linkage disequilibrium 49

3) Study of the genetic structure of the *M. polymorpha* collection 50

II) Patterns of selective pressures on *M. polymorpha* ssp. *ruderalis*' genes 54

1) SNP distribution and allele frequency spectrum in *M. polymorpha* 54

2) Presentation of the neutrality test statistics used 56

a) General concept behind Tajima's *D*, Fay and Wu's *H* and Zeng's *E* 56

b) Considerations on methods for calculating the *D*, *H* and *E* statistics 59

3) Calculation of *D*, *H* and *E* statistics in three plant species 62

a) *D*, *H* and *E* calculation on *M. polymorpha* ssp. *ruderalis* SNPs 63
b) Statistics calculation in *A. thaliana* and *M. truncatula* 63
4) Evaluation of intraspecific selective pressures in *M. polymorpha* genes 65
a) SNP distribution in the light of the unfolded SNP dataset and diversity indicators 65
b) Genes under selective pressures in *M. polymorpha* 67
5) Investigation of interspecific comparisons of intraspecific selection signatures in land plant species 69
III) Concluding remarks on chapter 1 71

Chapter II: Genome-wide association studies to understand *M. polymorpha*'s adaptation to biotic and abiotic stresses

I) Genome environment association study in *M. polymorpha* 76
1) Preparation of the data and GEA implementation 76
a) Retrieval of the climatic data for *M. polymorpha* ssp. *ruderalis* accessions 76
b) GEA analysis 79
2) Details on the candidate regions found 80
a) General description of the results 80
b) Enrichment on the full list of candidates 81
c) Focus on some candidates 82
3) Concluding remark on the GEA 94
II) GWAS of the response of *M. polymorpha* to a fungal pathogen 96
1) Material and methods 96
a) Experimental design and data modelling 96
b) GWAS model implemented 100
2) Candidate regions 102
a) Description of the different candidate regions for the different phenotypes 102
b) Focus on the terpene synthases 113
III) Concluding remarks on the genome wide association studies (GWAS on the response to *C. nymphaeae* and GEA) 120

Chapter III: The first pangenome in a bryophyte: construction and exploration

I) Building *M. polymorpha's* pangenome: a trial-and-error process 124
1) First trial: the map to pang strategy 124
2) Second trial: preliminary k-mer cleaning 128
3) Final idea: a gene-oriented pangenome 130
De novo assembly and structural annotation of accessions' sequences 131
Evaluation of genes presence absence variation through orthogroup inference 134
II) *M. polymorpha's* gene-based pangenome: the outcomes 135

1) Pangenome of the Marchantia complex 135
2) Pangenome of *M. polymopha ssp ruderalis* 138
a) General description 138
b) *M. polymorpha ssp ruderalis* core genome 142
c) *M. polymorpha* ssp. *ruderalis* accessory genome 144
d) Comparison of *M. polymorpha ssp ruderalis* core and accessory genome expression 149
4) Presence-absence-based genome-wide association studies 151
a) PAV-GWAS analysis on *M. polymorpha* symptoms of *C. nymphaeae* infection 153
b) PAV-GEA analysis on climatic data 155
III) Concluding remarks on the pangenome 162
Discussion and perspectives
I) *Marchantia polymorpha* main adaptive candidate genes 168
a) The terpene synthases 169
b) The peroxidases 170
c) The NLRs 171
d) The pathogenesis related (PR) proteins 172
e) The lectins 173
f) The lipoxygenases 173
II) *Marchantia polymorpha* population genomics characteristics 174
III) Perspectives on interspecific comparisons 179
Carbon footprint of the PhD project 184
List of figures 188
List of tables 191
Appendix 194
Bibliography 204

Introduction

I) The importance of genetic diversity in evolution

Evolution is the pivotal mechanism that shapes the essence of life, as it may be considered that the capacity to adapt coupled with heredity is the turning point of life emergence (Kunnev, 2020). The genetic material, its transfer and its modification are at the basis of biological evolution, functioning in a feedback loop with its phenotypic counterpart: variations in the genetic material causes variations in phenotypes, that may be inherited or not, depending on the advantage they confer to the organism in a given environment. Nevertheless, it should be noted that sometimes the selection on traits relaxes, and some phenotypic variations can also be transmitted without being strongly selected for. This potentially happens for quantitative traits, that display a continuous variation pattern and are governed by multiple genes with each variant under very weak selection, or when ecological niches are vacant, as it happens after a mass extinction (Kimura, 1991). Relaxation of the selective pressures allow new variants to emerge without being counter selected, variants that represent a fertile ground for new evolutionary progress. Even under selective constraints, most genetic variants are actually hypothesised to exist by chance and not because of a specific selection. This postulate comes from the neutral theory of molecular evolution, which posits that after emerging by random mutations, most variants are subjected to random fluctuation of allele frequency in the population, building a reservoir of diversity. Either way, the astonishing capacity of organisms to adapt to their surroundings always relies on their genomic diversity.

This genetic diversity arises through diverse random mechanisms (Figure 1). A DNA sequence can be impacted by spontaneous mutations occurring via DNA replication errors, mutagen-induced DNA damage, chromosomal rearrangement, or even polyploidization. This can lead to a variety of variants: single nucleotide polymorphisms (SNPs), insertions and deletions (InDels), copy number variation (CNV), inversions, translocations, tandem repeats. Recombination is another source of genetic diversity since it allows to shuffle the genetic information by exchanging portions of homologous chromosomes. The occurrence of recombination is not uniformly distributed throughout the genome, some regions of organisms' chromosomes being "recombination hotspots". Genetic diversity can also be brought by the transfer of genetic material between organisms that are not related, that is horizontal gene transfer. This mechanism is well known in prokaryotic genomes, leading for instance to antibiotic resistance in bacteria, but also brings diversity into eukaryotic genomes, to a higher extent than what was

previously though (Soucy et al., 2015). Introduction of a new gene in an organism provides selection with a novel combination of gene sequences to act upon.

All these sequence modifications can then be passed on, or lost, depending on demographic and selective forces. The latter tend to have a stronger effect on variants affecting the fitness of the organism. Neutral variants, that are variants in which the multiple alleles perform equally in terms of survival and reproduction of the individuals possessing them, are not subjected to strong selective forces, and their frequencies are therefore mainly influenced by demography, strongly linked with population size.

Genetic drift, that is the random fluctuation of neutral allele frequency across generations, is enhanced when the effective size of the population (*Ne*) is low. This is why populations undergoing bottleneck often undergo drastic changes in allele frequency. According to the neutral theory of molecular evolution, this random fluctuation is the main force exerted on most variants (Kimura, 1991). Population structure, gene flow between populations can also influence allele frequency. Population structure can lead to increased diversity between populations (through genetic drift and selection), the migration of an individual between two differentiated populations allows introducing new alleles in the recipient population, and balancing out genetic differentiation.

Mating systems can also impact the genetic diversity, for instance, selfing species have a lower proportion of heterozygosity and a lower effective recombination rate, and thus a smaller effective population size (D. Charlesworth & Wright, 2001). Assortative mating influences the genetic diversity: mating can be more or less frequent between specific individuals, compared to what is expected in random mating, leading to clusters of similar genotypes in groups of preferred mating partners (Wright, 1921).

As previously mentioned, for non-neutral variants, allele frequencies can be affected by selection forces (Figure 1). Selection is the non-random difference in reproductive success between organisms, partly due to differences in survival in a given environment. This difference in survival results from the interaction between the phenotype of the organism and the environment: some phenotypic traits confer survival and/or reproductive advantage in given conditions. The individuals that have these traits have a better fitness in this environment. This leads to the transmission of beneficial traits to the next generation. If maintained in this

environment, the frequency of these traits and the underlying genetic elements will increase along the generations. Natural selection works on the basis of genetic diversity: it favours some variants over others, leading to genetic diversity reduction, but at the same time genetic diversity powered by genomic forces (like mutation) creates new variants on which natural selection will act (Gregory, 2009).

There are three different types of selection signature, depending on the evolution of allele frequencies. If new mutations are counter-selected because they are deleterious compared to the alleles already present in the population, we will talk about purifying or background selection. On the contrary, if a new mutation is selected because it confers an advantage, the positive selection leads to a selective sweep. Finally, if diversity at a locus is favoured by the maintenance of several alleles, it is a case of balancing selection. When a locus is under selection, the other loci physically linked to it (close enough not to be separated by recombination) will also be affected, leading to characteristic signature of selection on a segment of chromosome, flanked by recombination events. For instance, background selection causes an excess of rare alleles in low frequency in the region (since it is always the original genomic segment passed to the progeny, and not the ones with new mutations). Selective sweeps can also cause an excess of rare since new mutations appear during, but also after, the phase where the haplotypes bearing the advantageous allele become quickly fixed in the population. Balancing selection, for its part, causes an excess of alleles in intermediate frequencies. For more details about these three type of selective forces, see the second part of this introduction and the second part of chapter 1.

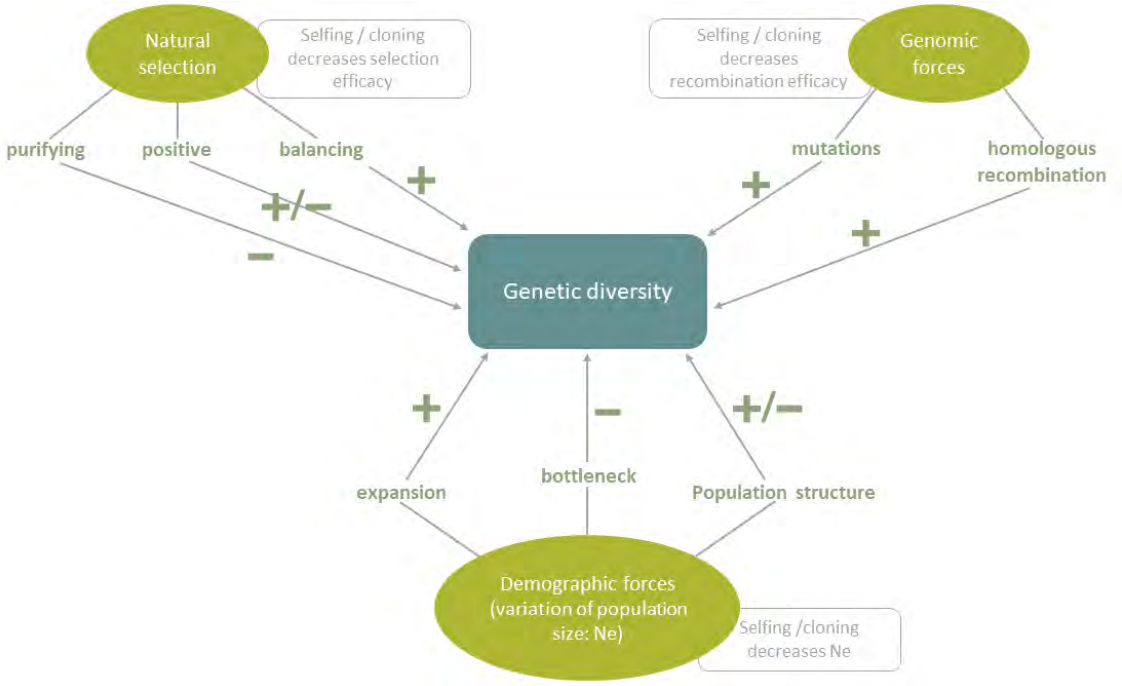


Figure 1: Summary of the genetic diversity drivers. Natural selection and genomic and demographic forces are the main factors that act on genetic diversity. The different events comprised in these main factors are highlighted and their effect on the genetic diversity is specified by "+" (increase of genetic diversity) and " –" (decrease of genetic diversity). The influence of the reproduction system on each main factor is specified in grey. Genetic diversity can be measured by nucleotide diversity, haplotype number, allelic frequencies, allele number… Adapted from Léa Boyrie's PhD manuscript.

II) The study of genetic diversity: history and tools

1) A short history of genetic diversity studies

The history of the study of adaptation, evolution and genetics is based on two independent currents of work. On one side, are the conclusions of Mendel on heredity. This Moravian monk is mainly known for his work on peas and the transmission of their phenotypic traits, in the second half of the 19th century. He spent seven years crossing different varieties of peas and following the evolution of some traits (shape of seeds, colour, flower colours, height of the plant) in the progeny. He quantified this evolution with mathematical tools, which was not common in biological studies at this time. This led to principles of biological heredity, called the Mendel's laws of inheritance. The law of dominance states that in heterozygous condition, the traits of some alleles, that are dominant will mask the effect of the recessive alleles. The law of segregation states that when gametes form, alleles are separated, and each gamete carries only one allele for each gene. Finally, the law of independent assortment states that the factors determining different traits are independently passed to the offspring (except for linked genes) (Monaghan & Corcos, 1984).

On the other side, is the work of Charles Darwin on evolution (completed by the work of some of his contemporaries like the explorer Alfred Russel Wallace), based notably on observations made during the *HMS Beagle's* expeditions. He theorized that organic beings vary in their traits, and that this diversity is a reservoir for some useful variation in certain conditions, that give individuals a "best chance of being preserved in the struggle of life" (Darwin, 2009). The Malthusianist idea of the struggle of life (competition for resources that can be food, sexual partners…), and the variation it acts upon constitute the fundamental bases of natural selection. Adaptations shaped by the natural selection accumulate and can lead to differences between local populations that results in the emergence of new species. These divergences can be due to geographic barriers, such as rivers, mountains (allopatric speciation) or genetic barriers, like temporal differences in reproduction period or polyploidy (sympatric speciation), that both prevent gene flow between the newly formed species. Speciation confers an advantage since the differences between newly formed species can allow them to colonize an expanding range of environments, and to reach ecological niches where they will be subjected to less competition. These ideas allowed to see living organisms as a continuum, rather than only partitioned species, and to change the focus of studies from specific "types" (specimen) to whole populations (Oldroyd, 1986).

The concepts laid down by Darwin allowed making sense of the intraspecific diversity and adaptation (microevolution), as well as the speciation processes and the inter specific diversity (macroevolution). He tried to explain heredity by a theory of pangenesis, in which each part of the body emits a « gemmule » that bears the instructions for the development of the corresponding organ (Oldroyd, 1986). It took around 60 years to bring together the infinitesimal variation model of Darwin with the discrete variations observed by Mendel, and to make the natural selection theory a genetic theory. The pangenesis theory was transformed by Hugo De Vries in a "pangene" theory, where traits are inherited in organisms by the transmission of particles called "pangenes" (a term that was later shortened to "genes"). To try understanding the underlying process, De Vries and other scientists studied the hybridization of plant species, unaware of Mendel's work, and obtained the same results, which helped bringing back interest on Mendel's results. Biometricians (supporting Darwin's theory) and Mendelians were in conflict, until R.A. Fisher connected the two theories, in the beginning of the 20th century, giving birth to the field of quantitative genetics. He understood that the infinitesimal variations

observed by Darwin actually resulted from the combined effect of multiple Mendelian factors (genes) (Berry & Browne, 2022). With Haldane and Wright, he is also considered as the founder of population genetics, having developed mathematical models to describe the genetic changes within populations over time. The three had some divergent views on the evolutionary processes. Wright focused mainly on the effect of stochastic factors on evolution, by studying genetic drift, whereas Haldane and Fisher though natural selection had the biggest influence. Fisher emphasized the role of natural selection in driving evolution, while Haldane focused more on the role of mutations in this process, by generating genetic diversity (Dronamraju, 2015). The studies of variation took a turn when molecular tools were developed, allowing to take a glimpse at the bases of variation, rather than relying only on phenotypic variants. These tools were first developed on protein sequences, with technics such as electrophoresis that allowed to detect variations in the sequence (B. Charlesworth & Charlesworth, 2017). Then studies on DNA were made possible, with the development of markers highlighting differences in sequences. The first ones were the Restriction Fragment Length Polymorphisms, that identified differences in the length of DNA fragments cut by restriction enzymes. Following this, microsatellites became the most popular genetic markers, based on short repetitive DNA regions located throughout the genome. These simple sequence repeats have varying numbers of repeats between individuals and are prone to mutations (processes of insertion/deletion of repeats), making them good markers to observe polymorphisms patterns (Grover & Sharma, 2016). With the development of sequencing technologies (*i.e.* Sanger), sequence targeted markers were developed. Single Nucleotide Polymorphisms (SNPs), that are specific single nucleotide differences, are one of the most used markers. The density of markers detected increased with the advances in Next-Generation Sequencing NGS technologies, allowing to reveal a diversity of polymorphisms in coding and non-coding regions. The detection of SNPs is based on the mapping of individuals' sequences on a reference genome, but with the continuous advances in long-read sequencing and assembly technologies, comparisons of whole genome sequences can now be carried out, allowing to discover rare variants and several types of structural variations (for more on that see the third part of this introduction chapter, on pangenome).

2) SNP-based methods to study population evolution

With the information of whole genome polymorphisms, the history of a species and of its adaptation can be investigated, through different tools. Most of these approaches belong to a field called population genetics, that aims at providing understanding of how evolutionary change occurs, by observing temporal and spatial variations of genotype frequencies.

a) Studies of demographic history

Intraspecific population genomics analyses enable the identification of diverse genomic patterns that reflect the history of a species, like population structure and population size evolution.

The measure of the fixation index (F_{ST}) allows quantifying the extent of genetic drift between distinct populations, and therefore their differentiation (Wright, 1965). This method needs a predefined idea of the outline of the different populations. If this information is not known, it is possible to define the number of distinct populations and their boundaries, by model-based and non-model-based methods. The non-model-based methods are based on the multivariate analysis of the matrix of markers (such as SNPs) in each individual. This can be done by PCA (implemented in PLINK for instance (Purcell et al., 2007)), or with more sophisticated approaches such as the discriminant analysis of principal components (DAPC). It is based on a round of dimension reduction by PCA, but then allows formally discriminating the different clusters of individuals, with a discriminant analysis that determines clusters by focusing on between group variability (Jombart et al., 2010).

Inference of clusters of individuals can also be performed via model-based methods that relies on statistical approaches, such as Bayesian clustering that is based on the definition of a probabilistic model describing the genetic profiles of individuals, and the use of Bayesian inference to update parameters of the model to fit the data, and simultaneously assign individuals to populations and determine populations allele frequency. Well-known softwares such as STRUCTURE (Pritchard et al., 2000), ADMIXTURE (Alexander et al., 2009), or FastStructure (Raj et al., 2014) rely on this strategy. Another approach to study population structure can be the phylogenetic approach. SNP matrix can be used to construct phylogenetic trees detailing the relationships between individuals. This can be used in concert with other structure analysis to have a detailed understanding of the relationships between different clusters of individuals.

From the 1990s to the present day, many population genetics methods have been developed to estimate the parameters of the demographic history of populations (estimation and variation of population size, split time between populations, migration rates, etc.) using multilocus genetic data. These methods benefited from a major breakthrough in the early 1980s in the understanding of the link between demography, genealogical processes and the genetic diversity of populations, thanks to the coalescence theory (Kingman, 2000). Significant advances have since been made in the field of genealogical exploration for likelihood inference, and hence parameter inference, with the importance sampling and Markov chain Monte Carlo (MCMC) methods (Beerli & Felsenstein, 1999; Nielsen, 1997; Stephens & Donnelly, 2000). The large panel of population genomic methods that infer the demographic history of populations will not be presented in more details, as they are not directly within the scope of this thesis.

Most of the tools mentioned tools can be used to infer the recent evolution of population and species, and can therefore be used in conservation genetics, where population genomics allow understanding the spatial and temporal evolution of populations, to be able to take actions to protect the species and guarantee its genetic diversity. Indeed, genetic diversity gives the organisms a reservoir for adaptation, and is often threatened by the genetic drift present in endangered populations with small population size. For instance, population genomics tools have allowed to evaluate the mutational load in two populations of kakapo, a flightless parrot, endemic to New Zealand and in critical danger of extinction since the 70's (Dussex et al., 2021). This will give a guideline of action to the breeding programs, in order to limit inbreeding depression in this small population size species.

b) Studies of selection footprints

Polymorphisms do not only bear the global species or populations' history, but also the evolutionary history of the loci. Genomes can therefore be scanned to detect selective patterns on loci. The frequency-based method relies on the signature that different type of selection leaves on the genome, and that can be differentiated by the distribution of alleles on polymorphic sites like SNPs: the site frequency spectrum (SFS). At different polymorphic sites of a DNA sequence, a reduction of sequence diversity can reflect background selection in the case of the prevalence of the ancestral allele, or a recent directional selection (selective sweep) if the derived allele is predominant. If the sites in the genomic region have an excess of intermediate frequencies, it is the sign of a balancing selection (Weigel & Nordborg, 2015). A

variety of test detecting these signals exist, like Tajima's *D*, that allow discriminating between balancing and background selection or selective sweep (Tajima, 1989), Fay and Wu's *H* that detects ongoing selective sweeps (Fay & Wu, 2000) or Zeng's *E* that distinguishes background selection from completed selective sweeps (Zeng et al., 2006). Detection of tracts of linked polymorphisms not broken down by recombination, can also be used, indicating fairly recent or ongoing positive selection (Siol et al., 2010). This haplotype fixation is the outcome of the hitchhiking effect where neutral polymorphisms spread in the population, carried along by their physical link (i.e. linkage disequilibrium, LD) with the adaptive polymorphisms.

Another method to detect selection relies on the comparison of polymorphisms' frequencies between populations, differences being indicative of local adaptations. For instance, Wright's F_{ST} technics compare the level of polymorphisms between individuals inside the populations and between individuals in different populations, to determine if there has been a higher level of differentiation between the two populations at the locus, compared to a genome-wide average (Beaumont, 2005; Bonhomme et al., 2010). Other methods of macroevolutionary comparison are based on protein-coding DNA sequences, such as the d_N/d_S test that compares the rate of non-synonymous substitutions (d_N) to the rate of synonymous substitutions (d_S), which is a neutral baseline that allows evaluating if there is an excess of non-synonymous substitutions. If it is the case ($d_N / d_S > 1$), it can be the sign of positive selection acting on amino acid residues, while a $d_N/d_S < 1$ would reflect purifying selection (Vitti et al., 2013).

Applied to plants, this can allow identifying genomic regions of crops genomes influenced by the domestication process, or targets of adaptation in wild populations and therefore to bridge the gap between the genotype and phenotype of adaptive traits.

c) Examples of applications in plants

Genetic signatures of selection and inference of demographic history are complementary in understanding the evolution of wild or domesticated plant populations. Population structure and demography are specifically important to take into account when carrying out SNP-based studies, since they can be confounding factors. For instance, small population size can result in higher Linkage Disequilibrium LD, and a fake signal of positive selection, or in random modifications of the SFS that can mimic the effect of natural selection (Siol et al., 2010). Studying individuals from distinct populations as a homogenous unit can lead to false signals of

balancing selection with intermediate frequencies of alleles, that are actually differentiated between the populations.

In crops, population genetics methods allowed determining the origin and migration of some major crops, as well as the genes that were selected during the process. For instance, with a genomic dataset of domesticated and wild African rice, researchers could pinpoint the origin of domestication of this species to a reduced area in Northern Mali. They could also uncover the genomic regions commonly selected between two domestications events of rice (in Asia and in Africa), like the loss of function of the PROG1 gene, responsible for the erect plant architecture, and other regions specific to one of the domestication events (Cubry et al., 2018). Crossing studies on different species, the co-occurrence of the domestication of cereals and legumes in various cradles of agriculture was brought to light, consistent with their complementary nutritional benefits (cereals serve as a carbohydrate source, and legumes as a protein source) (B. Song et al., 2023).

The regions found particularly impacted by the domestication process can be either genes of interest selected during the process, as the PROG1 gene, genes that were lost with the reduction of diversity due to the bottleneck of domestication. Some functions of interest can also have developed in wild species but not in cultivated ones, such as the SUB1 gene cluster that allows rice to withstand complete submergence and that appeared in some cultivated varieties via introgression (Fukao et al., 2009). These types of genomic regions are good targets for crop improvement, the first to optimise desirable agronomic traits, and the second and third to (re-)introduce them in the crop, since they are often involved in the response to biotic or abiotic stresses. This can be done by crossing wild relatives or traditional landraces with crops or by direct gene editing of the region of interest (Bohra et al., 2022).

Therefore, genomic studies on crops can shed light into the emergence of human societies, allow understanding past climatic and environmental changes, and help responding to future challenges.

In wild plant species, these studies provide the potential to infer the historical pattern of spreading of the species. For instance, in *Arabidopsis thaliana*, studies on SNPs allowed understanding that the modern population results from different relict populations that got isolated during the last glacial period. Only one of these populations underwent a strong and

quick expansion, leading to most modern *A. thaliana* individuals. Analysis of population differentiation allowed detecting a difference in flowering time regulators with some other relict populations. This could have been one of the advantages that facilitated the spread of this population (Alonso-Blanco et al., 2016). Genomic studies can also pinpoint the drivers of evolutionary successes of some species. A study on oaks found that their ability to rapidly migrate and adapt to new conditions during interglacial periods, coupled with a high rate of ecological divergence between oak clades and strong gene flow between them guaranteed their past evolutionary success (Kremer & Hipp, 2020).

d) Quantitative genetics: crossing genetic polymorphisms with phenotypes (or other traits)

Loci linked with adaptation can also be detected via their correlation with phenotypic variability, even though the association of a locus with a phenotype does not always mean it has an adaptive role, and has to be confirmed with complementary selection/fitness analyses. Compared to methods only based on genomic data, this type of method enables to directly know which adaptive phenotypic trait is influenced by the genomic loci considered adaptive.

Studies of genotype and phenotype association began with linkage mapping in population derived from crosses between individuals with contrasted phenotypes and genotypes. Haplotypes coming from each parent are broken down in the progeny of the crosses, and traced via their physical association with markers, like microsatellite. By observing the phenotype in the crossed population, regions coming from one of the parents could be correlated to a specific trait shared by the parent and some of its progeny, allowing Quantitative Trait Locus (QTL) mapping.

Nevertheless, these approaches are slow and labour intensive, and are often supplanted by Genome Wide Association Studies (GWAS), facilitated with the increasing ease of sequencing and SNP calling. The concept behind GWAS is to reveal correlations between genetic polymorphisms and quantitative traits variability via statistical models, taking advantage of historical recombination events in natural populations which show higher genetic variability than cross-derived populations. The most frequently used polymorphisms are SNPs, determined from a natural population, which provides a higher resolution than with genetic mapping on markers present in crosses' offspring. Initially, methods for GWAS were developed

in the field of human disease genetics, the first successful GWAS being the study of the genetic risk factors of myocardial infarction (Ozaki et al., 2002). Plant science had abundant genetic resources and a great interest in understanding the genetic architecture underlying complex traits, such as those related to agriculture (yield improvement, pest resistance, …). This interest allowed these strategies to be implemented quickly after their development (Tibbs Cortes et al., 2021), the first study outside of human medical genetics was therefore an investigation on GWAS feasibility in plants performed on *A. thaliana* (Aranzana et al., 2005).

The GWAS tests the association between each SNP and the trait of interest, via a defined model (linear model and derivatives) that allows to quantify the effect of each loci on the phenotype and to compute a p-value that indicates the significance of this effect on the phenotype. Since the numbers of tests performed are huge (one per SNP), multiple testing correction is performed to limit the false positive rate (Tibbs Cortes et al., 2021).

In the models, population structuration and relatedness between individuals must be accounted for, since they can generate false associations between the genotype and the phenotype. Indeed, in the case of two distinct populations with distinct phenotypic values, an association could be observed between this phenotype and a lot of loci, but these loci can bear different alleles (or alleles frequencies) simply because of populations genetic differentiation, without any causal link with the difference in phenotype. Individuals' population assignation and relatedness matrix are therefore considered in the mixed linear model (MLM) of the GWAS (Figure 2).

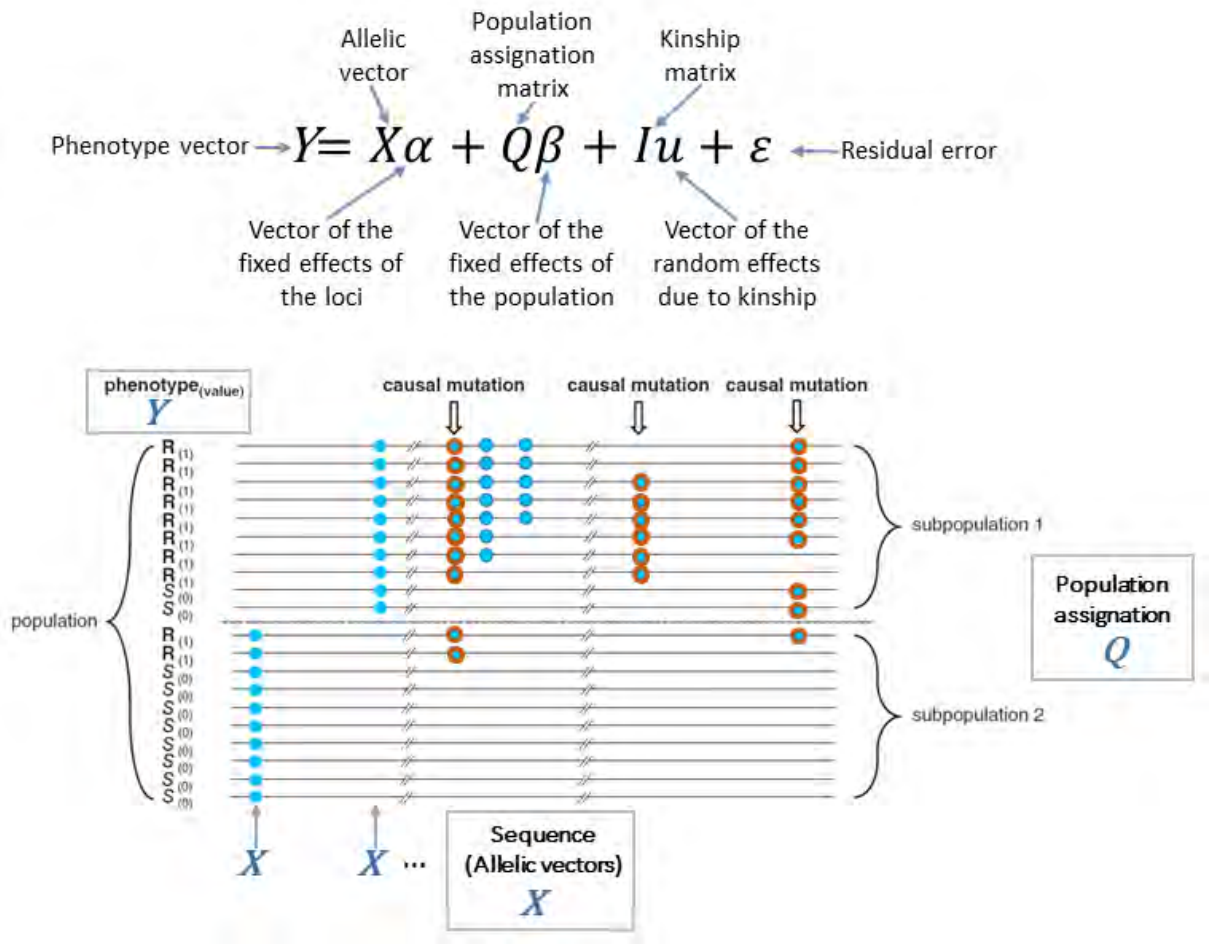


Figure 2: Illustration of the mixed linear model that can be used to test the association of a SNP with a quantitative phenotype of interest (Kang et al., 2010; Z. Zhang et al., 2010; X. Zhou & Stephens, 2012). On the upper panel is the formula of the linear mixed model that can be used, that explains the distribution of the phenotype among individuals (Y) with the vector of the alleles/genotypes present at the tested locus for each individual (X). To avoid the confounding effect of population structure, relatedness between individuals (I) and population assignation (Q) are taken into account. The parameter α that quantifies the effect of the SNP on the phenotype is the one that is then tested to see if it is significant. The bottom panel illustrates the confounding effect of population structure, by representing the allelic distribution along a genomic region (blue circles represent mutated alleles) in different accessions. These accessions all display a phenotypic value (R or S, the phenotype is binary for simplification). The two first SNPs of the genomic sequence seem correlated with the phenotype, but their distribution is actually linked to the allelic differentiation of the two population at this site. These false positive will not be considered as causal mutations with the mixed linear model correction for population structure and relatedness. Bottom panel figure adapted from Bonhomme and Jacquet 2020 ***(Bonhomme & Jacquet, 2020)****.*

GWAS have limits, the main ones being issues related to population structure and identification of candidates that only explain a small part of the heritability of complex traits. This missing heritability can be due to low power of the analysis to detect rare causal alleles, even the ones with large phenotypic effect, especially in low sample size analysis (Manolio et al., 2009). Some causal markers can be missed because they don't reach the significance threshold, especially with the stringent multiple testing corrections (Tam et al., 2019). Another reason is the

important role of polymorphisms larger than SNPs, like structural variation (SV), in the variability of plant phenotypes. So far GWAS have mostly been using SNP data, and missed out this aspect of genetic polymorphisms, but as analyses of long-reads assemblies expands, SV will be increasingly considered in GWAS (J.-M. Song et al., 2020; Y. Sun et al., 2022).

To increase the chances of detection of some variants, a local score approach can be used (Bonhomme et al., 2019). It allows amplifying the association signal by exploiting linkage disequilibrium between SNPs in a region containing a quantitative trait locus (QTL). This is done via the research of a maximum of a Lindley process: all the significant SNPs, that are below a given p-value threshold of $10^{-\xi}$, with the ξ parameter often being equal to 2 or 3, have a positive association score. This positive association score is cumulated for all the adjacent SNPs, thus increasing the Lindley curve (that is initially equal to zero), until a trail of non-significant SNPs is encountered, making the Lindley process decrease back to zero. This increases the power of detection of small effect QTLs, while major effect QTLs can still be identified (Figure 3).

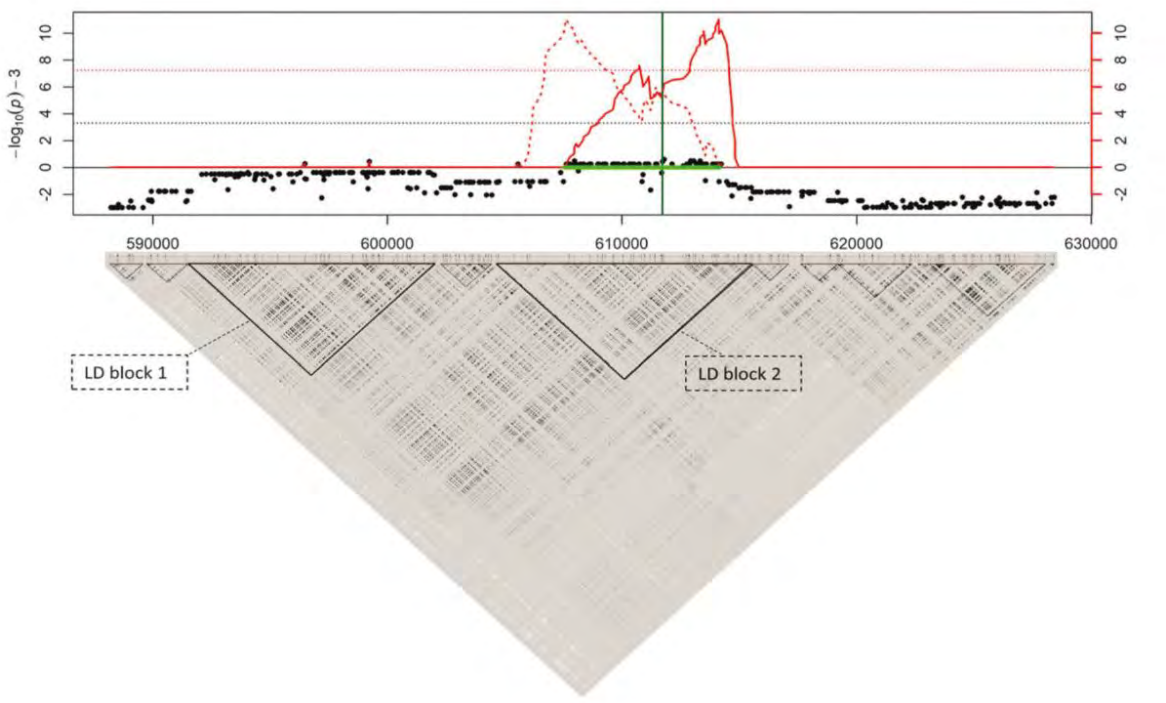


Figure 3: Illustration of the local score approach. Because of linkage disequilibrium (LD block 2), the statistic signal surrounding the QTL (green vertical line) is more significant than further up or down the genomic region. The association signal in this LD block is cumulated by the local score process, because it is above the p-value threshold (red horizontal line). The local score is represented by the red curves (the solid curve represent the cumulative process going from left to right, the dotted curve the process from right to left). The LD block 1 is not detected by the local score because its p-values are below the threshold. Figure extracted from Bonhomme et al. 2019 ***(Bonhomme et al., 2019)****.*

The traits studied via genome wide association can be phenotypic (resistance/susceptibility to a given pathogen, plant architecture, flowering time or production of some metabolites like flavonoids), or characteristics of the environment of the plant (climate of origin, altitude…) (Alseekh et al., 2021). In the latter case, the association analysis can also be termed genome-environment association (GEA) and implies that the species is locally adapted to its environment (Lasky et al., 2023).

Association studies have mainly been performed in crops or in model species, even though there are recent developments in a broader range of non-model species, with the advance of sequencing technologies. In crops, these studies have allowed to discover loci associated with traits of agronomic interest (concerning yield, plant architecture, stress tolerance or even metabolic diversity with mass-spectrometry-based analysis), that can be used to accelerate the breeding process (Alseekh et al., 2021). There is also an increasing work on orphan crops or wild relatives of crops, in order to uncover new genes that may have been lost in crops during the domestication process, or gained in the wild relatives, and offer resilience to abiotic and biotic stresses (Cortés et al., 2022).

In non-crops species, the main focus of studies is of course the model organism *A. thaliana*, so much so, that a catalogue of all hundreds of GWAS performed on this species has been created (Togninalli et al., 2018). This profusion of studies has allowed discovering the general genetic bases of *A. thaliana's* adaptation to climatic conditions, such as development time, flowering, or photosynthetic processes, but also specific candidate genes with previously unknown functions like AGG3, linked to cold temperatures (Ferrero-Serrano & Assmann, 2019; Hancock et al., 2011). Biotic stresses were also investigated, leading to the identification of 55 QTL of disease resistance in the past 9 years (Demirjian et al., 2023). Other model plants, like *M. truncatula* have also been well studied, leading to similar conclusions, on the importance of flowering time for instance (Burgarella et al., 2016; Yoder et al., 2013).

The genetic bases of adaptation are mostly found to be highly polygenic, with transcription factors, signalling proteins and epigenetic mechanisms often identified as candidate genes in GWAS (Frachon et al., 2018). This reveals the complexity of these mechanisms, governed by large gene networks. The extant of such gene networks might be so important that it is hypothesized that nearly all genes expressed in a specific condition could impact the trait, even in a very low magnitude. This hypothesis is called the "omnigenic" model (Boyle et al., 2017).

Nevertheless, some traits are governed by single strong effect genes, like the RPM1 NBS-LRR receptors (NLR) in *A. thaliana* that confers resistance to *Pseudomonas synringae* expressing specific avirulence factors (Boyes et al., 1998).

3) To complement the SNPs: Pangenomic studies

Structural variations are common in plants, generated by TE activity (Della Coletta et al., 2021), non-allelic homologous recombination (NAHR) (Parks et al., 2015), genetic introgression/horizontal gene transfer (HGT) (Ma et al., 2022), and biased gene loss (fractionation) in polyploid plants (F. Cheng et al., 2018). With the sequencing of whole genomes becoming more and more common, it has appeared that a single reference genome was not a comprehensive representation of the genomic diversity in a species (X. Yang et al., 2019). SNPs and short insertion deletion are widespread and easily detected polymorphisms across the genome. Structural variations (SV), for their part, are hardly detected by the mapping of short reads from individuals on a reference genome. Structural variations can be insertions, deletions, inversions, translocations (of 50 bp or more), copy number variation (CNV) or presence absence variations (PAV) of genomic regions. Since these large polymorphisms can vary a lot from individual to individual, the use of a single reference genome can bias the understanding of a species genomic content. A study in maize proved that, depending on the reference used to call SNPs, the candidates from genome-wide association studies vary, especially since they can be located in PAV regions (Gage et al., 2019). Similarly, in soybean the resistance to soybean cyst nematode is mediated by a 31 kb locus (Rhg1) with CNV between accessions. The multiple copies in the resistant accessions allow higher expression of the genes in the locus (Cook et al., 2012).

Following this rationale, the first pangenome has been created in 2005 by Tettelin and collaborators (Tettelin et al., 2005). They discovered that the total genomic content of the species *Streptococcus agalactiae* could be described by ordering genes in different categories: the core genes, shared by all strains and the dispensable genes, shared by only some strains. The rare genes present only in a few strains can also represent a category. The size of the pangenome of the bacteria depending on the number of strains added was also modelled, introducing the concept of the closed or open pangenome (Figure 4). The pangenome of this bacterium was open, meaning that the proportion of dispensable genes was so important that each strain added to the pangenome increased the total number of gene in the pangenome.

By opposition, a closed pangenome saturates after a given number of individuals added, because there are few individual-specific genes. Open pangenomes are often found in bacterium, depending on their lifestyle and the amount of HGT occurring in the species, whereas in eukaryotes the pangenomes are mostly closed (Golicz et al., 2020). With these concepts of core/dispensable genome and open/close pangenome, this work set the bases for all the pangenomic studies that followed.

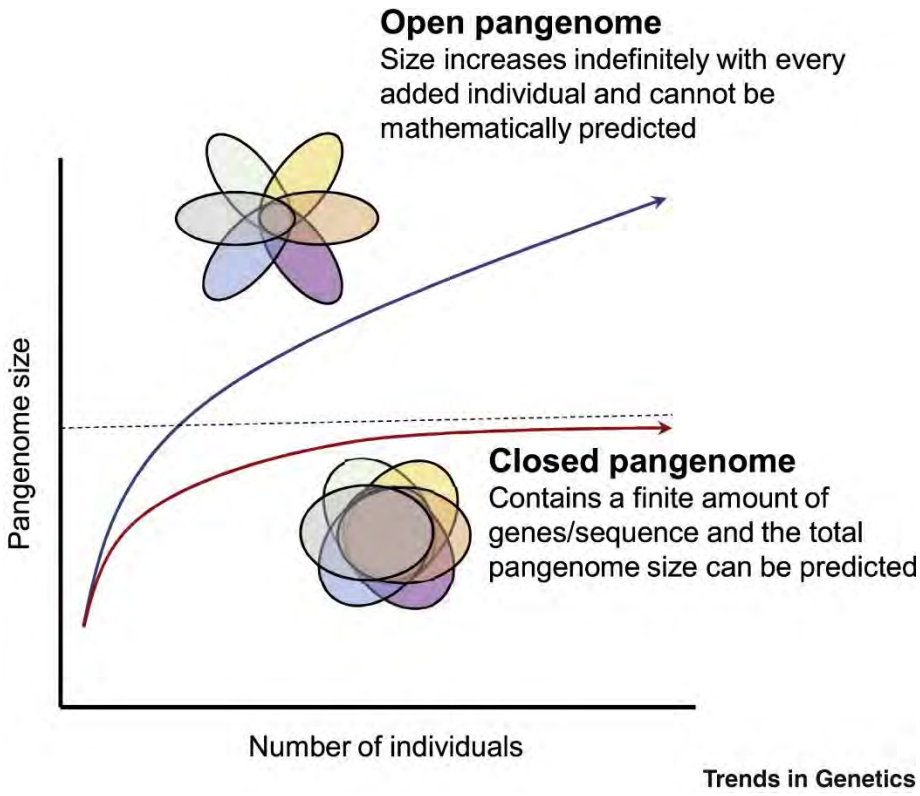


Figure 4: Representation of the concepts of open and closed pangenome, and core and accessory pangenomes. The two Venn diagrams represent pangenomes, with each oval representing gene (or genomic region) content from an individual. The circle in the middle represents the genes shared by all the individuals: the core genome, while the part of the oval on the outskirts of the Venn diagram represents the accessory genome, shared by few accessions. When the amount of shared genes is quite large compared to the size of the pangenome, the pangenome is closed: if another individual is added it will not increase greatly the number of sequences contained in the pangenome. In this case, the pangenome size can be predicted by modelling its saturating curve. On the opposite, when the core genome size is way smaller than the accessory genome, each new individual added will increase the pangenome size. Figure extracted from Golicz et al. 2020 ***(Golicz et al., 2020)****.*

Different strategies exist to construct a pangenome, depending on subject of interest and the type of data available (Figure 5). The "*de novo*" method was the first one created, consisting in whole genome assembly of each individual, followed by the comparison of the genomic contents. This method is more easily deployed on long-reads sequencing. To treat short-read sequences, the "map to pan" approach was developed. It consists in the iterative mapping of

accessions on a reference genome, followed by the iterative assembly of the non-mapping sequences, to create a pangenome reference composed of the initial reference genome and of the additional assembly (Bayer et al., 2020). Later on, the graph-based approach was developed, where a variation graph is constructed to represent the conserved and variable genomic zones between individuals. This type of graph is composed of vertices representing the common sequences and edges representing the variable regions (bubbles on the graph). This last type of pangenome offers the most complete and compact representation of the genomic diversity among individuals and can be used as a reference to map new individuals with increased accuracy, and even call SNPs (Baaijens et al., 2022). The construction of graph pangenomes can be based on multiple genome alignments (as the pggb software (pangenome graph builder) does (Garrison et al., 2023)) or local alignments of sequences to a reference genome backbone (as the minigraph toolkit does (H. Li et al., 2020)).

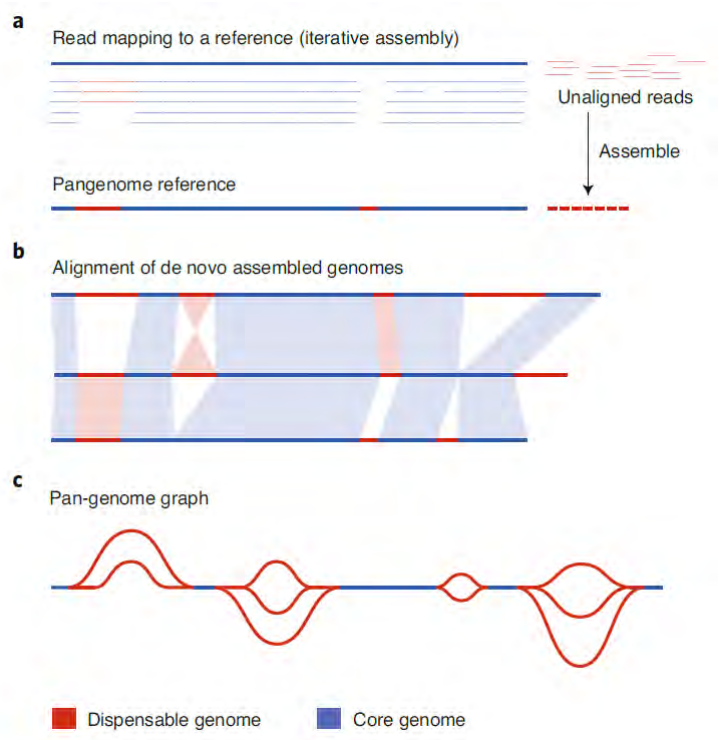


Figure 5: Different pangenome construction approaches. Are illustrated here the map to pan approach (a), the de novo method (b) and the graph pangenome (c). The blue elements represent the core genome, shared by all accessions, whereas the red elements represent the dispensable (or accessory) genome, only present in some accessions. In the map to pan approach the reads of each accession are mapped and the ones that do not map are assembled as part of the accessory genome. It should be noted that the reference genome itself contains accessory regions, on which some accessions does not map (cf the red regions in the reference part of the pangenome). The de novo assembly approach (b) gives a bit more information about big structural variations like inversions (cf the red "X" in the figure). The graph pangenome (c) is a compact representation that can represent some of the

structural variations (like the varying sizes of accessory regions (cf the different red edges). Figure extracted from Bayer et al. 2020 ***(Bayer et al., 2020)****.*

The choice of a strategy to construct a pangenome depends on the data available and the information wanted. The study of large structural variations requires a *de novo* approach or a graph approach with long reads and good assemblies, whereas the study of small SV and gene PAV can be run on short read data through a map to pan approach but could lack the information of their position in the genome.

As the variety of strategies shows, the pangenome is not limited anymore to the study of genes, as it was for the study of Tettelin and collaborators (Tettelin et al., 2005). The pangenome can therefore be described as the complete and non-redundant inventory of all the sequences of various individuals of a given species (or from another taxonomic level, such as genus, even though it is not as common). These sequences can be coding sequences (genes), but can also represent all the genomic sequences, taking into account variations in large fragments of the genome (structural variations).

Regarding the coding sequences, the gene PAV can be found by various strategies. With the mapping information, a common strategy is to evaluate the coverage of each accession's reads on each gene, in order to determine if the gene is present or not in the accession (for instance with SGSGeneLoss (Golicz et al., 2015), used for the lupin pangenome (Hufnagel et al., 2021)). With assemblies, synteny or orthogroup determination can be used to compare the sequences and link the orthologous genes (for instance, OrthoMCL has been used for the construction of the soybean pangenome (Y. Liu et al., 2020) and of the sorghum pangenome (Tao et al., 2021)).

The idea of a plant pangenome emerged in 2007 (Morgante et al., 2007) from the observation that TE were prevalent in plants, which caused dramatic differences in genomic content between individuals. This was followed (distantly) by the first plant pangenome on the wild relative of cultivated soybean, build on 7 de novo assembled accessions (Y. Li et al., 2014). Since then, pangenomics have been used on over 30 plant species, with a specific focus on crops (Bayer et al., 2020). The main conclusions of these studies were that the plants have a variable but non-negligible number of accessory genes, that are often involved in resistance to biotic or abiotic stresses (signalling components, immunity genes…). For instance, Nucleotide binding domain Leucine rich Repeat proteins (NLR) tend to often be contained in the accessory genome of plants, which led to the concept of the pan NLRome: a repertoire of all the NLR present in a

species (Barragan & Weigel, 2021; Van de Weyer et al., 2019), that allows to detect PAV of a whole gene or of a gene domain, that can be causal of the resistance/susceptibility of a plant facing a pathogenic microorganism. The pangenome approach is therefore a great complement of the classical analysis on plant diversity, since some genomic regions/genes are missing in the reference assembly (Q. Zhao et al., 2018). In contrast, the core genome often contains housekeeping genes and in general genes that are essential for the fundamental functions of the organism, linked to growth, cellular metabolism, cellular life cycle, reproduction... (Golicz et al., 2020).

The pangenomic approaches have also allowed to study the signatures of domestication in some crops by comparing them to their wild relatives. Genes responding to biotic and abiotic stresses are often lost during the bottleneck caused by the domestication process (Yu 2019 sesame, Gao 2019 tomato). This can allow to target interesting genes lost during domestication and reintroduce them in the crop species, in order to make them more resilient to various stresses. These tools can also be used to pinpoint regions and genes that confer desirable agronomic traits, like the loci associated to fiber characteristics, discovered in cotton (J. Li et al., 2021).

Pangenomes are gradually becoming the new references on which classical genomic analysis can be performed. For instance, GWAS have been performed on SNPs called on pangenome graphs, or on different polymorphisms. For instance, phenotypes can be compared to gene PAV patterns (Y. Sun et al., 2022; Tao et al., 2021) or to SV (He et al., 2023) to detect causal genes or regions. This combination of strategies allows to partially rescue missing heritability observed in classical GWAS analysis (Y. Zhou et al., 2022).

Resequencing efforts to uncover population diversity have focused on flowering plants, with a special emphasis on crops (B. Song et al., 2023). These efforts allowed to understand these plants evolutionary history and find genes linked with traits of interest. But this focus on flowering plants creates gaps of knowledge in the phylogeny of land plants, be it entire clades poorly studied (lycophytes, monilophytes, bryophytes), or bias in the coverage of species diversity in some angiosperm families. For instance in the Rosidae clade, the Malvales or Brassicales are extensively studied, whereas the Picramniales (a small neotropical family) has

almost not been studied (Marks et al., 2021; B. Song et al., 2023). Some effort have already been made in some non-crop plants like *Amborella trichopoda* (H. Hu et al., 2022) or *Gingko biloba* (Y.-P. Zhao et al., 2019), but they often have way smaller samplings sizes than crops studies. To better understand plant genome evolution through comparative genomics, it is essential to address these gaps in the knowledge and promote studying efforts on non-crop plants.

III) Land plant diversity and adaptations

Land plants, or embryophytes, are the descendants of a freshwater green algae belonging to the charophyte group and began to colonise land around 500 MY ago (Harris et al., 2022). This terrestrialisation led to tremendous changes on the planet, that shaped the world as we know it today. Atmospheric free oxygen comes from photosynthesis, that began with cyanobacteria 3 billion years ago, and algae 1.5 billion years ago (Rensing, 2018). Oxygen levels had already jumped during the great oxidation event, 2.3 billion years ago, but the colonisation of land by plants increased organic carbon burial, which caused a dramatic increase in O_2 levels, leading to present levels of oxygen (21%) in about 70 MYA (Lenton et al., 2016). This plant-mediated change in O_2 levels led to a new steady state of the earth system, that allowed among other things, an expansion of life, and the terrestrial clades diversification, such as large and mobile animals. The first land plants also increased physical and chemical rock weathering, compared to abiotic conditions (Porada et al., 2016), leading to soil formation. The other form of terrestrialised algae, lichens, also play a part in rock weathering but most probably arrived on land after land plants, between 440 and 250 MYA (Nelsen et al., 2020). The conquest of land by plants is therefore one of the deepest geobiological transition in Earth's history that established the basis of the contemporary ecosystems as we know them.

Plants, as photoautotrophs, are one of the bases of the food chains. It is no big surprise that the history of the human civilisation is therefore inextricably linked to plants. Today, plant represent the main food source (with six crops providing 80% of the human population's calories: wheat, rice, corn, potatoes, sweet potatoes and manioc), but also provide our societies with materials, chemical compounds (Whitton, 2013). One important aspect of plant science is therefore to develop strategies to optimise agricultural production. Today, with the anthropic climate change threat, the agricultural question is more than ever a hot topic, because of its

impact on the environment (nitrogen fertilisation, greenhouse gas emissions, soil degradation…) as well as the impact of extreme climatic conditions on the food production yield and quality. Multiple areas of plant research therefore focus on finding solutions to build a more resilient and more environmentally friendly agriculture (Eckardt et al., 2023).

Understanding the biology of land plants and the evolutionary processes that shaped them is, thus, one of the keys to ensure a stable agricultural production, but also, from a broader perspective, a key to understand how they helped shape our world as we know it today.

1) The diversity of land plants

As mentioned just before, land plants have freshwater origins, and are therefore part of the streptophyta algae division, with their closest sister clade being the Zygnematophyceae (Figure 6). Freshwater habitats are subjected to occasional drying, and some of the adaptation for life out of water already evolved in the common ancestor of this lineage, later facilitating the long-term transition of land plants from aquatic environments to terrestrial habitats. After millions of years of evolution in environments with different degrees of water availability, the different clades adopted distinct strategies. Many streptophyte algae are semi terrestrial organisms that can exploit transient hydration, shallow waters and interstitial moisture, and are highly tolerant to desiccation, while land plants, as multicellular terrestrial organisms, are able to regulate their water loss and to extract moisture from the depths of the soil (Delwiche & Cooper, 2015). With these new abilities, land plants thrived, and diversified into around 400 000 species (Christenhusz & Byng, 2016), distributed into various clades. They comprise two main lineages, the vascular plants (or tracheophytes) with a sporophyte-dominant life cycle and the non-vascular plants (or bryophytes) that have a gametophyte-dominant, haploid, life cycle. These two lineages diverged rapidly after plant arrival on land: the crown bryophytes originate 500-473 MYA whereas the crown tracheophyte originate 452-447 MYA (Harris et al., 2022). The tracheophytes include seedless plants: lycophytes (with approx. 1 000 species) and the monilophytes (ferns and allies, with approx. 14 000 species), as well as seed plants: gymnosperms (approx. 980 species) and angiosperms (approx. 352 000 species) (Whitton, 2013) (Figure 6). Angiosperm represent 80% of the extant plant diversity, due to multiple diversification bursts that occurred mainly during the Cretaceous and Paleogene, and could have been facilitated, among other things, by multiple whole genome duplications (WGD) (Benton et al., 2022).

The bryophytes form a monophyletic group encompassing approximately 22 000 species divided in three lineages: mosses (with an estimated 15 000 species), liverworts (approx. 7500 species) and hornworts (approx. 250 species) (Söderström et al., 2016). Bryophyte therefore represent a non-negligible part of the land plant diversity and show a worldwide distribution. As bryophyte depend on water for sperm dispersal, they occur in humid habitats. Their capacity to grow on a variety of substrates, including bare rocks, makes them pioneering species of crucial ecological importance. Bryophyte are characterised by genome reduction compared to the other lineages of land plants and to their common ancestor, that had been caused by pervasive gene loss (Harris et al., 2022).

On the other hand, most land plant lineages underwent one or multiple whole genome duplications (WGD). These events create redundant genes with relaxed evolutionary constraints that can be coopted into phenotypic innovations and adaptability, the so-called evolution by gene duplication (Ohno, 1970). Multiple WGD have been identified along the evolutionary history of land plants (Figure 6), some of which are named, like the epsilon WGD that occurred in the common ancestor of angiosperms. Among bryophytes, mosses is the only lineage with reported whole-genome duplication events (Clark & Donoghue, 2018).

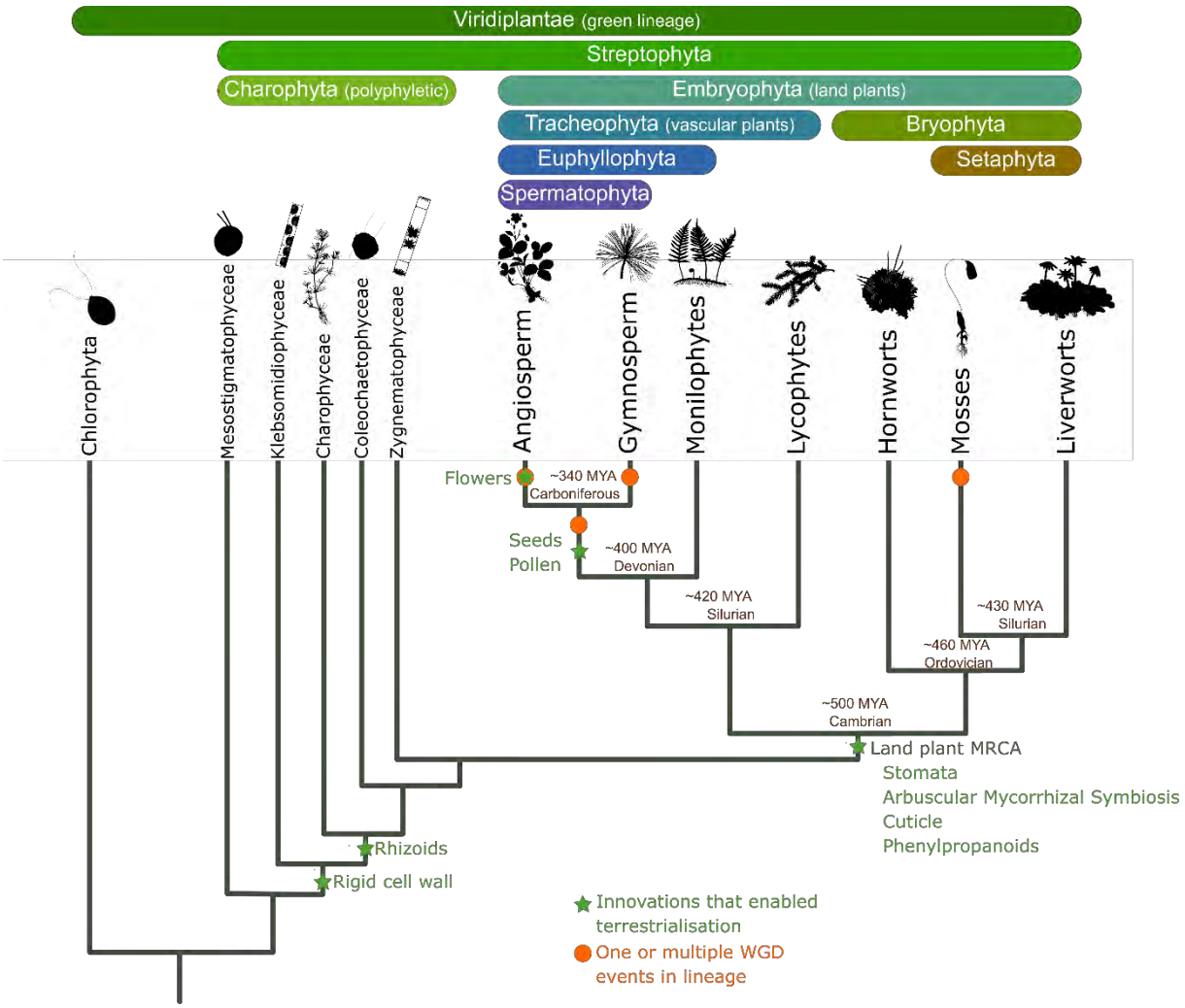


Figure 6: Phylogenetic relationships between clades of the green lineage. The topology of the tree is taken from (Puttick et al., 2018). The names of the different clades are indicated on top of the tree. Approximate times of divergence between clades are taken from (Harris et al., 2022). Whole genome duplications and some of the innovations that enabled the conquest of land by plants are indicated on the branches (according to Clark and Donoghue (Clark & Donoghue, 2018). Icons for the different plants are from Phylopic.org, with the icons for Zygnema and Klebsormidium from Phylopic/Matthew Crook (CC BY-SA 3.0), and the icons for Coleochaete and Chlamydomonas from Phylopic/Sergio A. Muñoz-Gómez (CC BY-SA 3.0).

2) Common "tools" to thrive out of water

The evolutionary journey of land plants has been marked with key innovations scattered all along the evolutionary tree, that permitted the successful and long-lasting colonisation of land. Adaptation from water to land and contact with the atmosphere required to adapt "terrestrial challenges" such as drought stress, UV radiation, new types of parasites, nutrient scarcity, and to develop new reproductive strategies less dependent on water (Delaux & Schornack, 2021; Rensing, 2018) (Figure 7). Once these challenges were overcome, plants benefited from the

unfiltered sun light and from the atmospheric CO_2 (with a concentration 50 times higher than the one in water) that enhanced photosynthetic processes (Delwiche & Cooper, 2015).

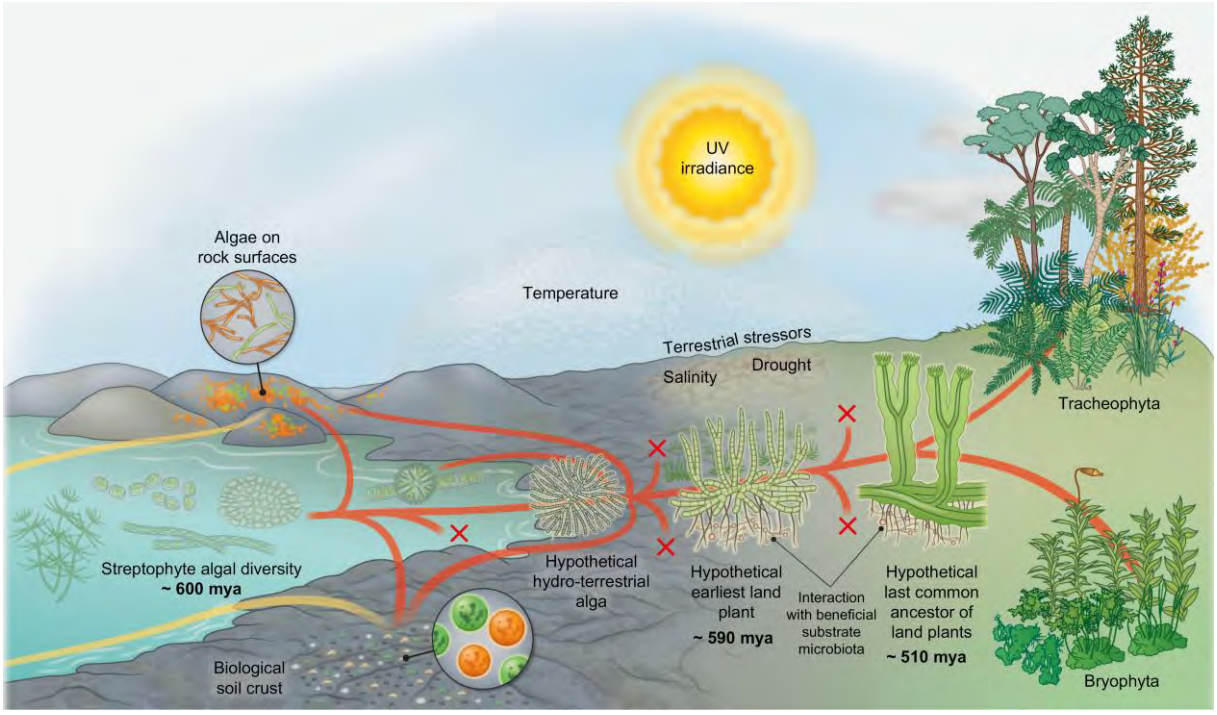


Figure 7: Illustration of a potential evolutionary scenario for the conquest of land by streptophytes. Streptophyte algae gave rise to the land plants (red lines) that evolved into two distinct lineages: tracheophyta and bryophyta. These algae also transitioned from water to land multiple times (convergence of yellow and red lines). Both land plants and terrestrial algae had to face multiple terrestrial stressors (UV radiance, temperature, salinity, drought…). Resistance to these stressors is mediated by a mix of ancestral algal genes and new adaptive genes. Retrieved from Fürst-Jansen et al. 2020 (Fürst-Jansen et al., 2020).

Multiple adaptations helped in this process, developed in the common ancestor of land plants during the innovative burst that occurred with terrestrialisation, or even prior to that (Figure 7). Sturdy cell wall protected the plant cells from abiotic stresses such as drought, UV radiations or gravity, but it also facilitated the development of complex body plans and vascularisation. This innovation is also present in charophytes and allowed them to live on land before the emergence of land plants (Harholt et al., 2016). Another key innovation of land plants was the ability to conduct water in their bodies, in order to control their water supply in a new environment that lacked water. This was made possible by specialised cells, that can be xylem cells in vascular plants, or hydroid cells in some mosses. These hydroid cells were shown to be regulated by the same transcription factors (NAC) than the xylem cells, suggesting an ancestral origin of this water conducting strategy (B. Xu et al., 2014).

The development of the biosynthetic pathway for phenylpropanoids (specifically flavonoids), that are UV-B absorbing pigments allowed plants to withstand a greater quantity of UV-B rays on land. The basic elements of this "sunscreen" pathway evolved early in the diversification of streptophytes and were then enriched by gene duplication, retention and subfunctionalisation (Rensing, 2018). The core pathway enzymes are therefore conserved across land plants, but some lineage specific modifications exist in the final enzymatic steps (Bowman, 2022). This is often the case for land plants essential biochemical pathways. For instance, the hydrophobic cuticle that protects plants against environmental stress, notably desiccation, is composed of a cross linked cutin scaffold and covered with a mixture of cuticular waxes. The biosynthetic pathways associated with these elements is partially present in streptophyte algae, and definitely emerged in the common ancestor of embryophytes. Gene family expansions then occurred in seed plants and monilophytes leading to lineage specificities in cutin synthesis (Kong et al., 2020).

Mutualistic symbiotic interaction with fungi is another trait that emerged in the common ancestor of land plant and facilitated land colonisation. Fungi from the Glomeromycota group associate with the plant's roots, in tracheophytes, or plant body, in bryophytes, and grants access to inorganic nutrients at the same time as it gives a protection against biotic and abiotic stresses. This symbiosis is called the arbuscular mycorrhizal symbiosis (AMS) and is pervasive in land plants. Most plants share the common signalling gene toolset to establish this symbiosis, even the ones that lost the ability to engage in AMS, but shifted to another intracellular symbiosis (like the ericoid symbiosis) (Radhakrishnan et al., 2020; Rich et al., 2021). A non-negligible part of land plants has lost this symbiosis (around 10 to 20%), potentially due to pathogen pressures or relaxed selection in nutrient-rich environments, but this loss is only stable if the plant develops a compensating mechanism, like another symbiosis (for instance, orchid mycorrhiza) or an alternative strategy for nutrient scavenging (for instance, carnivory) (Werner et al., 2018).

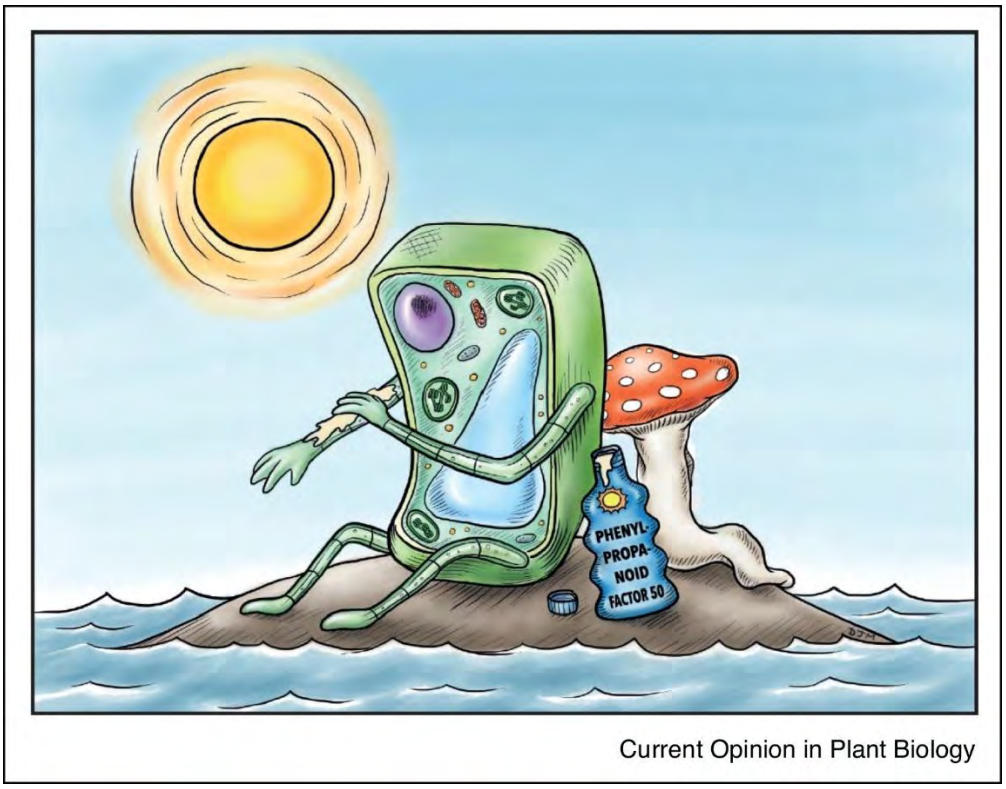


Figure 8: illustration of some of the key features that enabled plants to conquer land. The ones mentioned just before are the phenylpropanoid sunscreen, the rigid cell walls, and the symbiosis with fungi. Retrieved from Rensing 2018 (Rensing, 2018).

Land plants therefore share multiple traits that enabled the transition from an aquatic environment to the land, which was a consequent adaptive jump (Figure 8). This adaptive jump was made possible, genetically, by gene duplications or whole genome duplications and exaptation of older genes (as shown previously, most pathways' core gene sets were already present in streptophyte algae) (Donoghue et al., 2021). Another key element of genetic innovation was horizontal gene transfer (HGT) from other organisms, that brought additional genetic diversity for evolution to work on (Ma et al., 2022).

3) A diversity of clade-specific innovations associated with the diversification of land plants

Plant transition from water to land was a major change requiring multiple innovations to be long-lasting. After this major event, the different land-plants lineages continued to develop innovations that allowed them to thrive in different ecological niches.

For instance, flowers, seeds and pollen, that can come to mind as typical plant traits, are actually lineage specific innovations. Pollen and seeds appeared in spermatophytes (Figure 6)

and enabled water-independent reproduction, as well as a better resistance to environmental conditions than spores. Indeed, seeds are protected by their seed coat and can remain dormant until the conditions are favourable for germination (Linkies et al., 2010). Seeds are therefore used by spermatophytes to overcome (in a way) harsh environmental conditions, like a lack of water, by producing seeds that will germinate only when conditions are more suitable (Proctor & Tuba, 2002). Flowers allowed angiosperms to improve pollination efficiency by attracting animal vectors, especially insects (but also bird or bats) as pollination helpers (S. Hu et al., 2008), and by reducing the size of the gametophytes (pollen grain and ovule) that offered efficient protection to these haploid phases (Benton et al., 2022).

Adaptation to the irregular water supply on land has not only been achieved with pollen and seeds, but also during the vegetative state, with two distinct strategies existing in land plants: homoihydry and poikilohydry. The first one has been adopted by tracheophytes and is based on internal water transport from the soil to the leaves, with regulation of the loss of water in these organs (with stomata, a waterproof cuticle…). The second strategy is used by bryophytes and consists in a passive equilibration of cells' water content, with the temporary cessation of metabolism when water is missing. This "ultimate drought evading mechanism" has also been adopted by some vascular plants, with at least 10 independent gains of this trait (Proctor & Tuba, 2002). Both mechanisms guarantee the availability of water during photosynthesis, either by transporting it to the photosynthetic organs, or by suspending photosynthesis when water is not accessible. The weaknesses of the homoihydric strategy are the dependence on constant availability of water in the soil and the limitations of water loss regulation. Indeed, there is a balance between CO_2 uptake necessary for photosynthesis and the water loss the plant can afford. Similarly, under heat conditions, evaporative cooling might be needed, leading to extensive water loss. The poikilohydric strategy weaknesses are the limited environmental range poikilohydric species can cover and their dependence on external factors. The two main lineages of land plant therefore have different ways of coping with the same major environmental constraint.

Among the bryophyte-specific adaptations, gemma cups can be cited. Gemma cups have been observed in liverworts, and are cup-shaped structures that contains clonal progeny called gemmae (Kato et al., 2020). These clones are then disseminated in the environment via a splash

cup dispersal mechanism (water carries the gemmae away from the plant), allowing the plant to rapidly expand and colonise its surroundings.

These examples show the diversity of traits and adaptations that exist in land plants. To better understand their distribution and evolutionary history in land plants, comparisons between species are used. These comparisons can be made on phenotypic traits, but these ones can be misleading, since convergent evolution makes it difficult to distinguish ancestrally inherited traits and convergently evolved ones. Genetic studies unravelling the evolutionary history of genes underlying specific traits are therefore the key to understand land plant evolution. To be performed, such studies require extensive data on the genome of diverse plants.

IV) Perks of studying a model bryophyte: *Marchantia polymorpha*

1) The importance of studying bryophytes (and, more generally, other understudied plant lineages)

Angiosperms are the most diversified clade of land plants and represent a primary resource of food, fibres, building material, pharmaceuticals for modern human societies. As a result, plants sciences have focused on flowering plants, and specifically crops (Marks et al., 2023). The knowledge on other plant lineages such as ferns, gymnosperms or non-vascular plants (bryophytes) is still lagging behind, and does not represent the richness of plants biological diversity. Nevertheless, there is a growing interest in these species and new model plants emerge in different clades, helped by the development of new technologies (RNAi, CRISPR/Cas9) (Chang et al., 2016). In a genomic point of view, the development of sequencing technologies has allowed expanding the range of species with sequenced genomes and to begin bridging this gap. Concerning bryophytes, some representative genomes are now available for the three main clades: four hornworts (two strains of *Anthoceros agrestis*, *Anthoceros punctatus* (F.-W. Li et al., 2020), and *Anthoceros angustus* (Zhang 2020)), six mosses (*Ceratodon purpureus* (Carey et al., 2021), *Physcomytrium patens* (Lang et al., 2018; Rensing et al., 2008), *Sphagnum fallax* (Healey et al., 2023), *Pleurozium schreberi* (Pederson et al., 2019), *Fontinalis antipyretica* (Yu et al., 2020) and *Hypnum curvifolium* and *Entodon seductrix* (Yu et al., 2022)) and six liverworts (*M. inflexa* (Marks et al., 2019), *M. paleacea* (Radhakrishnan et al., 2020), *M. polymorpha ssp montivagans* and *M. polymorpha spp polymorpha* (Linde et al., 2020), *M.*

polymorpha ssp ruderalis (Bowman et al., 2017a) and *Lunularia cruciata* (Linde et al., 2021). However, the sequenced species often belong to the same clades, leading to a heterogenous coverage of bryophyte phylogenetic diversity so far.

The key phylogenetic position of bryophytes, that diverged from tracheophyte approximately 480 MYA (Harris et al., 2022), and the new genomic resources available from them, along with other land plants, led to the development of new fields of research. Evolutionary studies, like the evo-devo (evolutionary developmental biology) or evo-MPMI (evolutionary molecular plant-microbe interactions), aim at reconstructing trait evolution based on phylogenomic comparison of different species (Delaux et al., 2019). Indeed, by comparing bryophytes to other land plants, and to algae, it is possible to deduce which features were present in the most recent common ancestor (MRCA) of land plant and which are lineage specific. These methods provide a better understanding of the features that allowed the plant's terrestrialisation process, but also of the adaptations developed later on, in each lineage and allowed investigating most of the traits mentioned in the previous section. It is only under the light of comparison between different lineages that bryophytes are a great tool to expand our knowledge on the past traits of land plants, and not by assuming they are "basal" land plants bearing only ancestral traits. Indeed, contrary to the common conception, bryophytes are not living fossils that give a perfect image of the ancestral land plant. They are as "evolved" and as divergent to the MRCA of land plants as tracheophytes (McDaniel, 2021). Actually, it has been proven that concerning some features, angiosperms are more similar to the MRCA than bryophytes (Harris et al., 2020; Rich & Delaux, 2020). For instance, comparison of the gene contents between lineages has revealed that the MRCA probably had stomata (Harris et al., 2020) and a vascular system (Harris et al., 2022), while some modern day bryophyte don't. This is due to bryophyte-specific reduction of gene content monitoring these processes. Thus, their study also allows uncovering bryophyte synapomorphies: morphological and physiological adaptations that evolved only in these lineages, such as the gemma cups cited previously.

This is exemplified, for instance, by the growing interest in bryophytes' metabolites. Indeed, even though bryophyte have little nutritional value, they produce molecules with very interesting properties (Asakawa & Ludwiczuk, 2018). For instance, some bryophytes produce capsaicin and piperine (hot tasting substances), as well as vitamins (B_2 and E), that make them potential candidates for food additive production. Non-vascular plants have medicinal

properties since their metabolites have antimicrobial, antifungal, antiviral, cytotoxic or muscle relaxing activities. *Marchantia polymorpha* itself was historically used as a diuretic in Europe, and produces Marchantin A, a muscle relaxant. Finally, bryophyte's volatile and odorant compounds can also be used in cosmetics. A better understanding of bryophytes' biology will allow to optimally harness their properties.

Although the interest in bryophytes took time to gain momentum, even though they represent a non-negligible part of the diversity of land plant, they are now increasingly studied, and count a few model plants. Among them, the moss *Physcomitrium "Physcomitrella" patens* was the first model to emerge, followed by the liverwort *Marchantia polymorpha,* and more recently by the hornwort *Anthoceros agrestis*. Here we will focus on the liverwort model: ***Marchantia polymorpha***.

2) *Marchantia polymorpha* as a model bryophyte

Among the bryophytes, the liverwort *Marchantia polymorpha* stands out as an historical model, described since the Greek antiquity, and represented in naturalistic illustration as soon as the 15th century (Bowman, 2016). *M. polymorpha* belongs to the Marchantiopsida class that encompasses the Sphaerocarpales order (bottle liverworts), the Blasiales order and the Marchantiales order (complex thallus liverworts, that have dorsal air chambers) that includes *M. polymorpha*. The two other classes of liverworts are the Jungermaniopsida (that comprises two orders: the Jungermanniales (leafy liverworts) and the Metzgeriales (simple thalloid liverworts)) and the Haplomitriopsida (the basal sister group, that comprise the Haplomitrium, Treubia and Apotreubia) (Figure 9).

Inside of the Marchantiales order, the Marchantia genus, that gets its name from the 17th century French botanists Jean and Nicolas Marchant, houses the two closely related model species: *M. polymorpha*, and the AMS (arbuscular mycorrhizal symbiosis) forming *M. paleacea* (Radhakrishnan et al., 2020). *M. polymorpha* consists of three subspecies, that have diverged 5-7 MYA (Villarreal A. et al., 2016), and that differ only slightly in morphology, but significantly in ecology: *M. polymorpha ssp montivagans* is found at higher altitudes and often grows in sunny wetlands forming deep mats of upright thalli, *M. polymorpha ssp polymorpha* also grows in natural damp habitats, but usually as thalli tightly appressed to rocks or clayey soil and finally, the most established genetic model system, *M. polymorpha ssp* ruderalis is a rapid coloniser that lives in human-altered environments (Shimamura, 2016). Reciprocal crosses between the

three subspecies are not always successful but there is evidence of gene flow between lineages, through hybridisation and introgression (Linde et al., 2020).

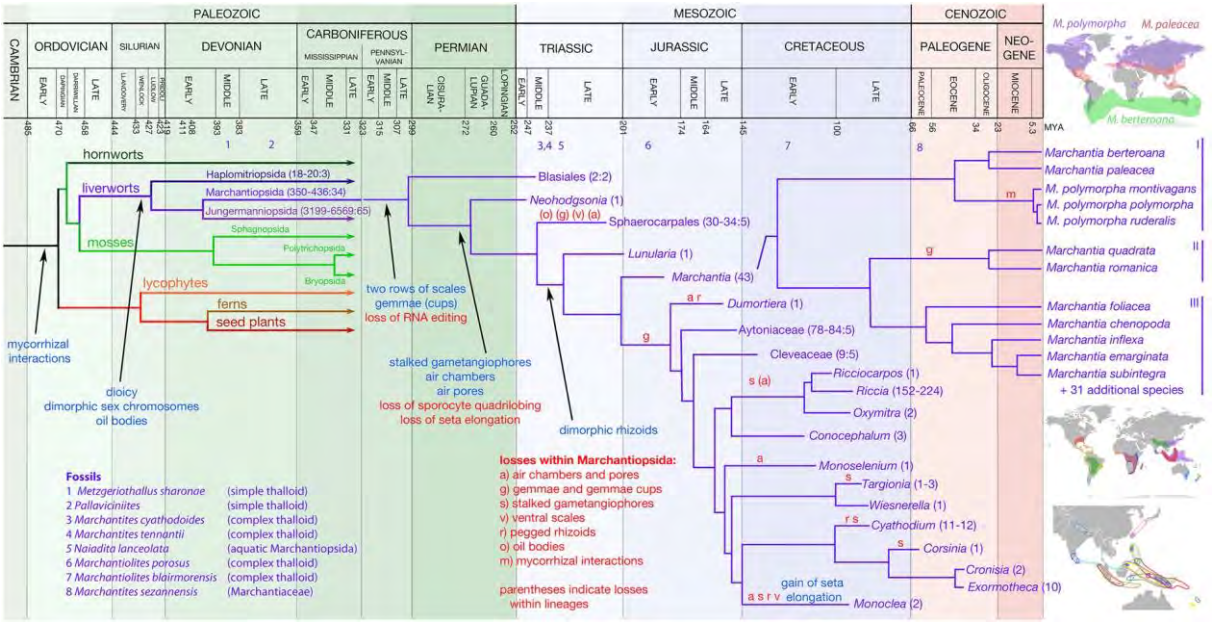


Figure 9: Phylogenetic context of Marchantia polymorpha, plotted along the geological timescale from (Bowman et al., 2022). Clade specific function gains and losses are represented in respectively in blue and red.

M. polymorpha possesses some traits that are specific to bryophytes, or even to more specific lineages, and that may render it quite peculiar for flowering plant aficionados.

Regarding its morphology, Marchantia is a complex-thalloid liverwort with a dorso-ventral morphology (Figure 10). On the dorsal side is found a photosynthetically active cell layer and specific organs like air chambers, intracellular spaces that enable gas exchanges similarly to stomata but with no regulation to open and close them, and gemma cups that are disc shaped propagules that allow clonal reproduction. On the ventral side can be found ventral scales and two types of unicellular filaments called rhizoids: the smooth rhizoids that mediate nutrient uptake and anchor the thallus to the substrate (similar to tracheophytes' root hair) and the pegged rhizoids that allow water transport (Shimamura, 2016). Between them is a storage region with parenchymal cells, and idioblast cells that contain oil bodies: organelle specific to liverworts that synthesise and store large quantities of terpenoids and aromatic compounds (Romani et al., 2022).

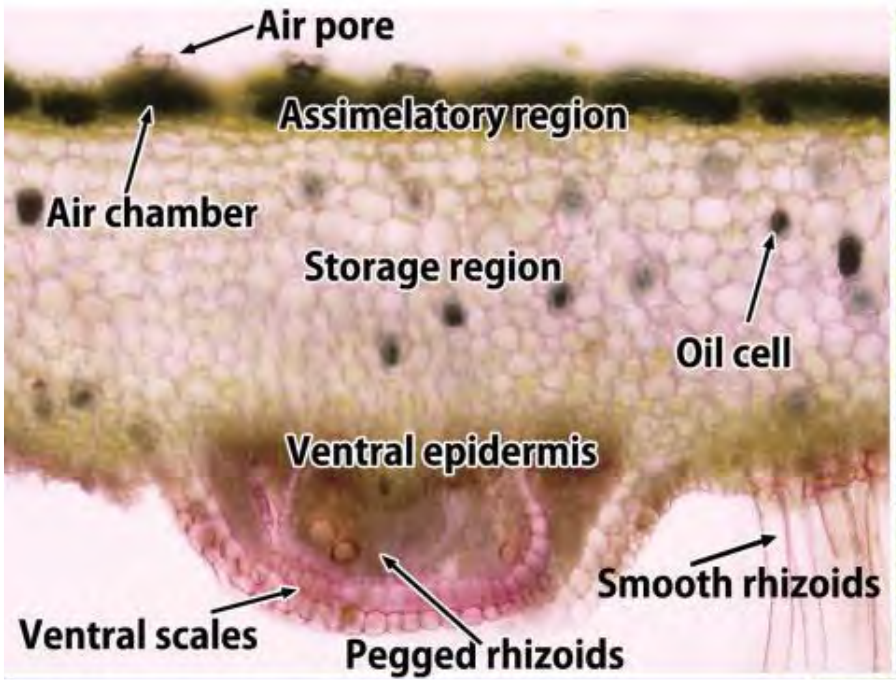


Figure 10: Transverse section of the thallus showing the dorsal and ventral characteristics of M. polymorpha. Retrieved from (Shimamura, 2016).

Marchantia has a gametophyte haploid dominant life cycle and is a dioicous species with sex chromosomally determined (V chromosome for the males, U for the females). During the sexual reproduction phase, the female gametophyte bears eggs on its reproductive structure (the archegoniophore), that are fertilised by sperm carried from the male antheridiophore by water, leading to the formation of a diploid sporophyte. Following meiosis, this sporophyte produces spores that give rise to new individuals (gametophytes) (Figure 11). *M. polymorpha* is also able to reproduce asexually via gemmae, contained dormant in gemma cups and which will develop into clonal thalli when dispersed by splashing water (Shimamura, 2016).

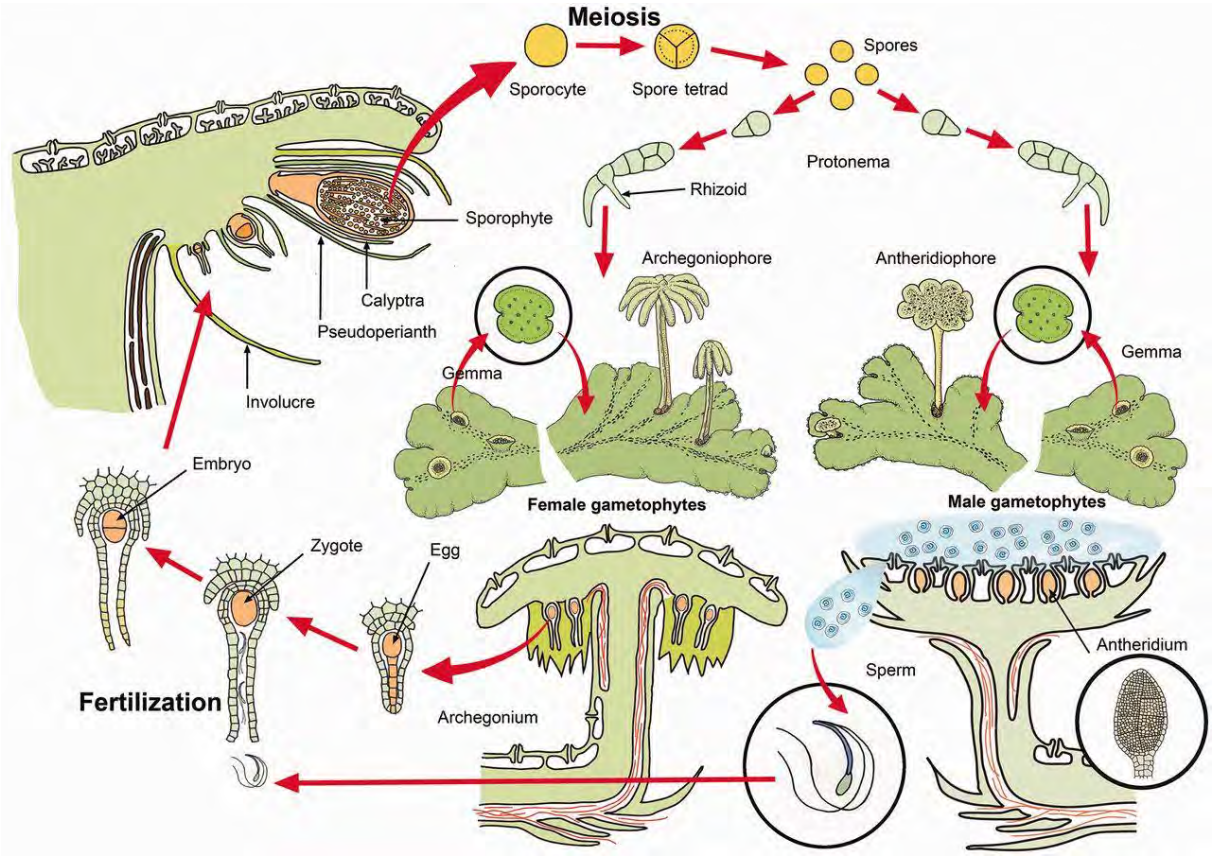


Figure 11: Life cycle of M. polymorpha from (Shimamura, 2016). This illustrates the sexual reproduction in M. polymorpha, with the alternance of gametophytic and sporophytic phases, and the asexual reproduction mediated by gemmae.

As previously mentioned, *M. polymorpha* is an historical model, on which a variety of studies have been conducted. It has been used to study reproduction in cryptogams and developmental polarity in plants (Bowman, 2016). *M. polymorpha* has also been a model plant of the beginning of the genomic era, with the sequencing of its chloroplastic genome (Ohyama et al., 1986) and the study of its sexual chromosomes (Yamato et al., 2007). More recently, this effort continued into the sequencing of the *M. polymorpha* genome (Bowman et al., 2017a), which enabled other omics studies like a gene expression atlas on the response of *M. polymorpha* to abiotic stresses (Tan et al., 2023), or a single cell expression study (Wang et al., 2023). Most of the genomic information on *M. polymorpha* is gathered on the Marpolbase website (https://marchantia.info/).

It also exists a great body of work on the interaction of *Marchantia polymorpha* with microorganisms. By comparing their microbiome to bacterial profile in the soil, it has been shown that *M. polymorpha* and *M. paleacea* were associated with classical plant-associated beneficial bacteria like *Rhizobia, Methylobacterium* and saprophytic species (Alcaraz et al.,

2018). Studies on the fungal endophytes of *M. polymorpha* have been carried by J. Nelson, revealing that most strains have little impact on the liverwort growth. Nevertheless, some species displayed growth promoting activity on Marchantia (*Nemania serpens, Biscogniauxia mediterranea or Colletotrichum truncatum*) although some of them were known as tracheophyte pathogens, while other species seemed pathogenic for the plant (*Xylaria cubensis, Colletotrichum sp 1*) (J. M. Nelson et al., 2018). Other studies have been carried on the infection of *Marchantia* with well-known phytopathogens, like the oomycete pathogen *Phytophtora palmivora that* colonises its air chambers (Carella et al., 2018, 2018), the bacterial pathogen *Pseudomonas syringae* (Gimenez-Ibanez et al., 2019) or the fungal pathogen *Fusarium oxysporum*, even though it is known to colonise the plant xylem in tracheophytes (Redkar et al., 2022). Those studies revealed the conservation of many plant immunity mechanisms in *M. polymorpha*: the presence of receptor kinases (BAK1, CERK1), the accumulation of salicylic acid (SA) leading to defence gene activation, the role of the pathogenesis related (PR) proteins, the mediation of the immune response by pathogen responsive transcription factors (WRKY, GRAS, bHLH) and the use of phenylpropanoid biochemical defences. This also allowed uncovering *M. polymorpha* specificities like the absence of FLS2 (FLAGELLIN-SENSITIVE 2) the flagellin receptor well-known in angiosperms, the presence of lineage-specific metabolites (like the anthocyanidin Riccionidin A) or the absence of Jasmonic acid (JA), functionally replaced by dn-OPDA. In Marchantia this JA precursor plays the same part in the antagonistic interaction with SA that balances the immune response depending on the type of pathogen (biotrophic or necrotrophic) (Gimenez-Ibanez et al., 2019; Matsui et al., 2020).

This existing body of knowledge makes *M. polymorpha* a convenient model for conducting biological studies. Its status as a model is also facilitated by its typical model organism characteristics (Cesarino et al., 2020):

-*Marchantia* is simple to cultivate thanks to its small size, its short life cycle (completed in about 2 months), and it is also easy to propagate either clonally with gemmae, or by crossing two individuals and retrieving the produced spores.

-Genetic transformation has been developed (Ishizaki et al., 2016).

-Genome editing methods such as CRISPR/Cas9 are available (Ishizaki et al., 2016; Sugano et al., 2018)

-The mostly haploid life cycle facilitates genetic and phenotypic analysis since there is no heterozygosity, and thus no dominance/recessivity issues. This can be a limitation when working on mutants of essentials genes for which mutations are lethal, but that can be circumvented by knockdown strategies like artificial micro-RNA gene silencing.

-*M. polymorpha* has a small nuclear genome of 220 Mbp with 8 autosomes and 1 sexual chromosome, with a bit less than 20 000 genes. The absence of ancestral whole genome duplication (WGD) in liverworts (Clark & Donoghue, 2018; Linde et al., 2023) grants this plant with a low functional redundancy that makes studying specific pathways easier. For instance, *M. polymorpha* possesses one gene of each category of transcription factor from the auxin signalling pathway known in angiosperms, as opposed to the multiple orthologs for each category of transcription factor that flowering plants possess (Kohchi et al., 2021). It should be noted that this is not the case in all bryophytes, since mosses underwent several rounds of WGD (Clark & Donoghue, 2018; B. Gao et al., 2022).

Marchantia polymorpha is therefore a well-established model plant (at least for a bryophyte) with a variety of resources to facilitate its study. But, at the same time, it still suffers from a lack of resources on specific aspects. One of them is the lack of a collection of accessions to study the intraspecific variability of traits and of genome in this species. This limitation is actually true for all bryophytes, more generally. Bridging this knowledge gap, to be able to deepen our knowledge on the genes and underlying mechanisms present in this bryophyte, and to study the variation of these elements in the collection to unravel adaptive mechanisms, was the main objective pursued during this PhD.

V) Thesis objectives

As explained previously, the study of land plants is still biased towards a few crops and model angiosperms, even though studies on non-angiosperm models or plants of economic interest are increasing. This PhD project therefore aims at expanding the genomic knowledge on the diversity of stress response mechanisms that exists in land plants. To do so, we built an intraspecific genetic diversity dataset in *M. polymorpha* and performed diverse analyses on it, with the objective of finding genes and gene families linked to the adaptation of this model liverwort to its environment. Findings made with these analyses were then crossed with the existing knowledge about other land plants (mostly angiosperms), in order to place our candidates within the context of global land plant evolution. The different approaches are summarised in Figure 12.

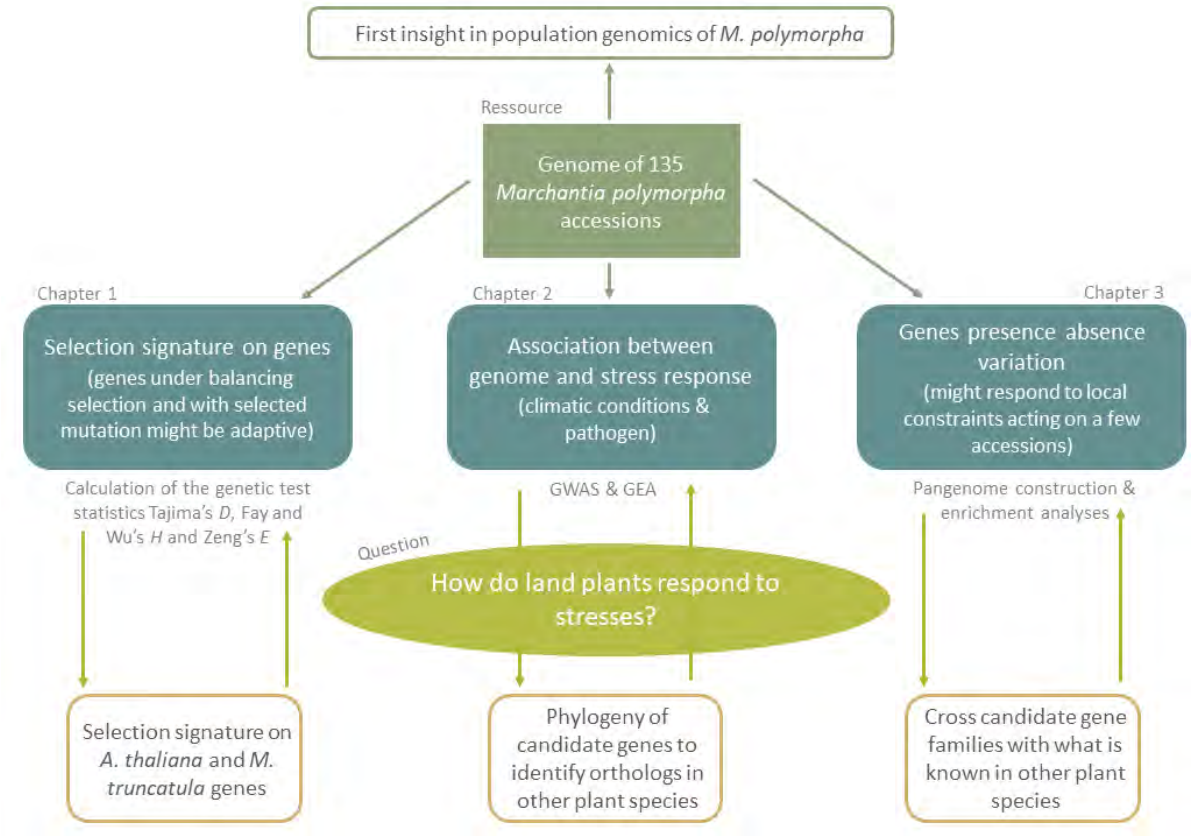


Figure 12: Summary of the PhD project, highlighting the main question and the approaches used to address it. The main methods used for each approach are specified under the blue boxes.

This manuscript is therefore articulated in three different chapters representing the three approaches used to uncover adaptation mechanisms in *Marchantia polymorpha*. The first chapter is based on the detection of selection signatures that can point to adaptive genomic

regions and the second chapter uses genome-wide association studies to link genomic variations to phenotypic variation. In this case, climatic data from the accessions' sampling sites and the response of *M. polymorpha* to a pathogen are considered. Finally, the third chapter aimed at exploring another aspect of genomic diversity by studying gene presence-absence through the construction of a gene-based pangenome. The study of accessory genes, that often results from local adaptations specific to a few accessions, provide other *M. polymorpha* adaptive candidate genes. In all chapters, different comparisons (detailed in the orange boxes of Figure 12) are made with data already available in other land plants, to better understand the evolutionary relevance of our gene family candidates.

This work should give new insights into the genomic bases of strategies evolved by land plants to adapt to their surroundings and highlight matters of interest for the bryophyte and for the land plant research communities.

Chapter I

Marchantia polymorpha diversity dataset and population genomics analyses

I) The *M. polymorpha* diversity dataset

1) Description of the collection

a) Collaborative sampling of a broad collection of accessions

The collection of *M. polymorpha* accession originates from a sampling effort from different research teams between 2013 and 2022 (Figure 13).

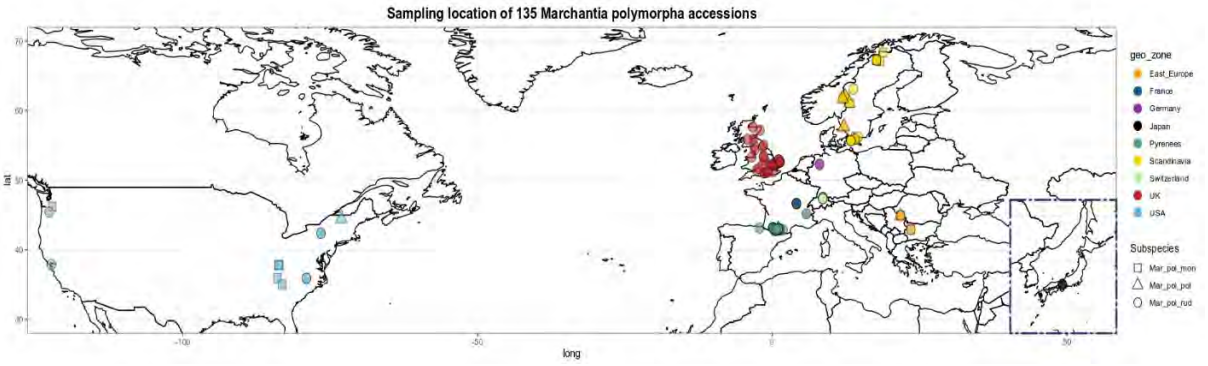


Figure 13: Sampling location for the accessions of M. polymorpha that constitute the collection. 127 accessions are represented here because the sampling location is unknown for 8 accessions. The dots representing the accessions are coloured depending on their sampling area, and the dots shape correspond to the subspecies of the accessions.

The accessions sampling in the USA was performed by Jessica Nelson (J. Nelson & Shaw, 2019), the one in Eastern Europe, Scandinavia and Switzerland by Peter Szovenyi (Linde et al., 2020). The majority of the accessions of the collection come from either the Pyrenees (36 accessions) due to a sampling effort carried out by P-M Delaux and F. Roux, or from the UK (35 accessions) with the participative science project "the Great Bristish Liverwort Hunt" (Hoey et al., 2023). To that are added the two BoGa accessions from the botanical garden of Osnabrück (S. Zachgo) and the Takaragaike-1 and 2 accessions (TAK1 and TAK2), that are the model accessions on which the reference genome of *M. polymorpha* was sequenced in 2017 (Bowman et al., 2017b). The collection sites showed a large diversity of micro-environments, from sidewalks to soil nearby ponds, and from sea level to an altitude of 1124 m (Figure 14, Table 1). The sampling effort was mainly focused on the *ruderalis* subspecies, leading to a sampling of 105 accessions from this subspecies, but the collection also contains 14 accessions from the *polymorpha* subspecies and 16 from the *montivagans* subspecies.

Figure 14: City marchantia, country marchantia. This panel displays some of the sampling sites from the Pyrennean sampling effort (left, pictures PMD) and from the great british liverwort hunt (right, pictures (Hoey et al., 2023)). This shows the diversity of microhabitats occupied by M. polymorpha, from plant community in natural soil, to pavement.

Table 1: General information for the 16 accessions from the montivagans subspecies, for the 14 accessions from the polymorpha subspecies, and for the 105 accessions from the ruderalis subspecies. The names of the accessions are indicated in the first column and some will be mentioned in the rest of the manuscript. For each accession, collection site and its GPS coordinates are indicated, as well as the inferred sex of the plant (based on a PCR on sex-specific markers and/or on the coverage of reads on the U and V sexual chromosomes).

id	Subspecies	site of collection	geo_zone	GPS_coord_decimal_deg	sex_consensus
NILSC12	Mar_pol_mon	Bjornliden(Sweden)	Scandinavia	62.04123066986471, 12.37654646947005	F
NILSC14	Mar_pol_mon	Bredvallen (Sweden)	Scandinavia	61.083758, 13.134897	
NILSC17	Mar_pol_mon	Engerdalsaetra (Norway)	Scandinavia	61.761465051440716, 11.956914809874396	F
NILSC18	Mar_pol_mon	Laitaure (Sweden)	Scandinavia	67.12621582431234, 18.254214197931006	F
NILSC20	Mar_pol_mon	Sarek (Sweden)	Scandinavia	67.28216133589784, 17.707824276137657	M
NILSC21	Mar_pol_mon	Sarek (Sweden)	Scandinavia	67.28216133589784, 17.707824276137657	
NILSC22	Mar_pol_mon	Sarek (Sweden)	Scandinavia	67.28216133589784, 17.707824276137657	
NILSC23	Mar_pol_mon	Abisko (Sweden)	Scandinavia	68.35767815033093, 18.81588620245602	M
NILSC24	Mar_pol_mon	Raven's combe Ogoya village Sophia district (E	East_Europe	42.911864, 23.513627	F
NILSC27	Mar_pol_mon	Waterfall at Vaioaga (Roumania)	East_Europe	44.91622654466679, 21.764009891497317	F
RES33	Mar_pol_mon	University of Tennessee, Knoxville, TN, USA	USA	35.95713, -83.92493	M
RES50	Mar_pol_mon	USA, KY, Powell County, Daniel Boone Nationa	USA	37.81891, -83.6796	M
RES51	Mar_pol_mon	USA, KY, Powell County, Daniel Boone Nationa	USA	37.8190333, -83.6788833	F
RES52	Mar_pol_mon	Chatooga River Gorge, North Carolina, USA	USA	35.01564, -83.12671	F
RES53	Mar_pol_mon	USA, WA, Skamania County, Mt, St, Helens pu	USA	46.24313, -122.16515	M
Bonheur	Mar_pol_mon	#N/A	NA	#N/A	
Caz-A	Mar_pol_pol	Cazavet (09)	Pyrenees	43.0016675690354, 1.041777521545995	F
NILSC1	Mar_pol_pol	Agusa	Scandinavia	55.7650040443062, 14.00107795323735	M
NILSC10	Mar_pol_pol	Lindome (Sweden)	Scandinavia	57.57374347842507, 12.07927319471927	
NILSC11	Mar_pol_pol	Bjornliden (Sweden)	Scandinavia	62.04123066986471, 12.37654646947005	
NILSC13	Mar_pol_pol	Bredvallen (Sweden)	Scandinavia	61.083758, 13.134897	F
NILSC16	Mar_pol_pol	Engerdalsaetra (Norway)	Scandinavia	61.761465051440716, 11.956914809874396	
NILSC2	Mar_pol_pol	Agusa	Scandinavia	55.7650040443062, 14.00107795323735	F
NILSC28	Mar_pol_pol	Waterfall at Vaioaga (Roumania)	East_Europe	44.91622654466679, 21.764009891497317	F
NILSC6	Mar_pol_pol	Horby (Sweden)	East_Europe	55.85754811617041, 13.649108585718611	
NILSC9	Mar_pol_pol	Jonsered (Sweden)	East_Europe	57.753277859734894, 12.185186659690952	F
Pra-B	Mar_pol_pol	Prat-Bonrepos (09)	Pyrenees	43.028667, 1.018083	
WRG	Mar_pol_pol	USA, VT, Chittenden county, Winooski River G	USA	44.48071, -73.11571	M
Maud	Mar_pol_pol	Cambridge	UK	52.188405, 0.117822	M
YC-A	Mar_pol_pol	#N/A	NA	#N/A	

Table continued

id	Subspecies	site of collection	geo_zone	GPS_coord_decimal_deg	sex_consensus
Ale-A	Mar_pol_rud	Aleu (09)	Pyrenees	42.89349317033232, 1.2682079283552368	
Ber-A	Mar_pol_rud	Bernac Debat (65)	Pyrenees	43.167583, 0.108667	F
Ber-B	Mar_pol_rud	Bernac Debat (65)	Pyrenees	43.167583, 0.108668	F
Ber-C	Mar_pol_rud	Bernac Debat (65)	Pyrenees	43.167583, 0.108668	
Ber-D	Mar_pol_rud	Bernac Debat (65)	Pyrenees	43.167583, 0.108669	F
Ber-F	Mar_pol_rud	Bernac Debat (65)	Pyrenees	43.167583, 0.108668	F
Bid-A	Mar_pol_rud	Bidania-Goiatz (Spain)	Pyrenees	43.140472, -2.159306	F
Bie-A	Mar_pol_rud	Biert (09)	Pyrenees	42.898306, 1.314972	F
Bie-B	Mar_pol_rud	Biert (09)	Pyrenees	42.898306, 1.314973	F
Bie-C	Mar_pol_rud	Biert (09)	Pyrenees	42.898306, 1.314973	
Bot-X	Mar_pol_rud	unknown	NA		M
Bul-A	Mar_pol_rud	Bulan (65)	Pyrenees	43.039806, 0.277361	M
Bul-B	Mar_pol_rud	Bulan (65)	Pyrenees	43.039806, 0.277361	M
Bul-C	Mar_pol_rud	Bulan (65)	Pyrenees	43.039806, 0.277361	
CA	Mar_pol_rud	Albany (near San Francisco) California USA	USA	37.89750637746934, -122.30761713258545	F
CAM1	Mar_pol_rud	Cambridge, near botanical garden (UK)	UK	52.194156837643504, 0.12375373851033176	M
CAM2	Mar_pol_rud	Cambridge, near botanical garden (UK)	UK	52.194156837643504, 0.12375373851033176	F
Cas-A	Mar_pol_rud	Castillon (09)	Pyrenees	42.920778, 1.032639	M
Cas-B	Mar_pol_rud	Castillon (09)	Pyrenees	42.919194, 1.031667	F
Cas-C	Mar_pol_rud	Castillon (09)	Pyrenees	42.920778, 1.032640	
Cas-D	Mar_pol_rud	Castillon (09)	Pyrenees	42.920917, 1.034833	M
Cas-E	Mar_pol_rud	Castillon (09)	Pyrenees	42.920806, 1.033861	
Dur-A	Mar_pol_rud	Durban sur Arize (09)	Pyrenees	43.021806, 1.350917	F
Esc-A	Mar_pol_rud	Esconnets (65)	Pyrenees	43.069167, 0.229333	M
Gen-A	Mar_pol_rud	Genos (65)	Pyrenees	42.811361, 0.403806	
Gen-B	Mar_pol_rud	Genos (65)	Pyrenees	42.811361, 0.403807	M
Hec-A	Mar_pol_rud	Heches (65)	Pyrenees	43.016333, 0.372000	
Hec-B	Mar_pol_rud	Heches (65)	Pyrenees	43.016333, 0.372000	F
Hec-C	Mar_pol_rud	Heches (65)	Pyrenees	43.016333, 0.372001	M
L2-BoGa	Mar_pol_rud	Bot Gard Osnabrück (All)	Germany	52.28088490939511, 8.028335178428128	F
L5-BoGa	Mar_pol_rud	Bot Gard Osnabrück (All)	Germany	52.28088490939511, 8.028335178428128	M
Lac-A	Mar_pol_rud	Lacourt (09)	Pyrenees	42.943750, 1.174528	F
LouisXIII	Mar_pol_rud		NA		F
Luc-A	Mar_pol_rud	Luchon (31)	Pyrenees	42.785722, 0.593667	F
Luc-B	Mar_pol_rud	Luchon (31)	Pyrenees	42.785722, 0.593668	M
Mns-A	Mar_pol_rud	Monsegur (09)	Pyrenees	42.870278, 1.833222	
Mnt-A	Mar_pol_rud	Montgaillard (65)	Pyrenees	43.122056, 0.106250	F
Mou-A	Mar_pol_rud	Moulis (09)	Pyrenees	42.960111, 1.088556	F
Mur-A	Mar_pol_rud	Muriannetes (38)	France	45.189972, 5.812278	
NILSC15	Mar_pol_rud	Bydalen (Sweden)	Scandinavia	63.100250709995684, 13.797574611794662	F
NILSC19	Mar_pol_rud	Sarek (Sweden)	Scandinavia	67.28216133589784, 17.707824276137657	F
NILSC25	Mar_pol_rud	Raven's combe Ogoya village Sophia district (E	East_Europe	42.911864, 23.513628	
NILSC26	Mar_pol_rud	Waterfall at Vaioaga (Roumania)	East_Europe	44.91622654466679, 21.764009891497317	F
NILSC3	Mar_pol_rud	Skrylle Stone quarry (Sweden)	Scandinavia	55.69385669149776, 13.358422922174048	
NILSC4	Mar_pol_rud	Roan Skane province (Sweden)	Scandinavia	56.2160736699062, 14.573313673489132	
NILSC5	Mar_pol_rud	Lyngsjo (Sweden)	Scandinavia	55.93503231543013, 14.071905993907915	F
NILSC7	Mar_pol_rud	Skrylle Christmas tree orchard (Sweden)	Scandinavia	55.693203771194725, 13.362100844812206	F
NILSC8	Mar_pol_rud	Skrylle Christmas tree orchard (Sweden)	Scandinavia	55.693203771194725, 13.362100844812206	M
Nor-A	Mar_pol_rud	Giles, Norwich, Norfolk (UK)	UK	52.633972, 1.296833	F
Nor-B	Mar_pol_rud	Cathedrale, Norwich, Norfolk (UK)	UK	52.631806, 1.300028	F
Nor-C	Mar_pol_rud	Ferry Lane 12, Norwich, Norfolk (UK)	UK	52.630750, 1.302806	F
Nor-D	Mar_pol_rud	Ferry Lane 9, Norwich, Norfolk (UK)	UK	52.630750, 1.302807	F
Nor-E	Mar_pol_rud	Almary Green, Norwich, Norfolk (UK)	UK	52.630694, 1.300278	M
Pra-A	Mar_pol_rud	Prat-Bonrepos (09)	Pyrenees	43.028944, 1.018194	F
Pra-C	Mar_pol_rud	Prat-Bonrepos (09)	Pyrenees	43.028639, 1.017583	M
RES29	Mar_pol_rud	Treman State Park, Ithaca, New York, USA	USA	42.39733115660473, -76.56144170578354	F
RES31	Mar_pol_rud	Irchel (Switzerland)	Switzerland	47.53801838538173, 8.606565972161894	
RES32	Mar_pol_rud	Irchel (Switzerland)	Switzerland	47.53801838538173, 8.606565972161894	F
RES34	Mar_pol_rud	Lewis & Clark College, Portland, Oregon, USA	USA	45.44962, -122.67003	F
RES35	Mar_pol_rud	North Carolina Botanical Garden, Chapel Hill, I	USA	35.89900, -79.034346	M
RES36	Mar_pol_rud	North Carolina Botanical Garden, Chapel Hill, I	USA	35.89900, -79.034347	F
RES37	Mar_pol_rud	Bot, Gard Zurich	Switzerland	47.35930327546087, 8.560506354701422	
RES38	Mar_pol_rud	Bot, Gard Zurich	Switzerland	47.35930327546087, 8.560506354701422	F
RES40	Mar_pol_rud	Bot, Gard Zurich	Switzerland	47.35930327546087, 8.560506354701422	
RES41	Mar_pol_rud	Bot, Gard Zurich	Switzerland	47.35930327546087, 8.560506354701422	
RES49	Mar_pol_rud	Lick Brook, Ithaca, New York, USA	USA	42.39997, -76.53836	F
Sal-A	Mar_pol_rud	Salechant (31)	Pyrenees	42.954833, 0.631417	F
TAK1	Mar_pol_rud	Takaragaike pond Kyoto Japan	Japan	35.05702548792268, 135.7811198646839	M
TAK2	Mar_pol_rud	Takaragaike pond Kyoto Japan	Japan	35.05702548792268, 135.7811198646839	F
Tou-A	Mar_pol_rud	Toulon-sur-Arroux (71) Fabrice Roux	France	46.650417, 4.113833	M
Tou-B	Mar_pol_rud	Toulon-sur-Arroux (71) Fabrice Roux	France	46.650417, 4.113834	M
Tou-C	Mar_pol_rud	Toulon-sur-Arroux (71) Fabrice Roux	France	46.650417, 4.113835	M
Tou-D	Mar_pol_rud	Toulon-sur-Arroux (71) Fabrice Roux	France	46.650417, 4.113836	M
Tou-E	Mar_pol_rud	Toulon-sur-Arroux (71) Fabrice Roux	France	46.650417, 4.113837	M

Table continued

id	Subspecies	site of collection	geo_zone	GPS_coord_decimal_deg	sex_consensus
Voe-A	Mar_pol_rud	Voewood (UK)	UK	52.914861, 1.119167	
Voe-B	Mar_pol_rud	Voewood (UK)	UK	52.914861, 1.119168	F
Aberdeen	Mar_pol_rud	Aberdeen	UK	57.173167, -2.083553	F
Beaufort	Mar_pol_rud	Winchester	UK	51.069026, -1.302273	F
Bristol	Mar_pol_rud	Bristol	UK	51.456017, -2.626775	F
Crowsnest	Mar_pol_rud	Nottingham	UK	52.895800, -1.243203	F
Dutchpot	Mar_pol_rud	Winchester	UK	51.046085, -1.358392	M
Field-1	Mar_pol_rud	Sheffield	UK	53.398312, -1.480008	F
Field-2	Mar_pol_rud	Sheffield	UK	53.398312, -1.480008	F
Gilesgate	Mar_pol_rud	Durham	UK	54.785081, -1.579859	M
Glo-1	Mar_pol_rud	Woodchester	UK	51.709502, -2.278406	M
Gormley	Mar_pol_rud	Cambridge	UK	52.188405, 0.117822	F
Grafton	Mar_pol_rud	Cambridge	UK	52.187632, 0.164367	F
Hopeman	Mar_pol_rud	Elgin	UK	57.651894, -3.309757	M
Jimw	Mar_pol_rud	Carlisle	UK	54.956240, -3.042268	M
Lakedist	Mar_pol_rud	Lake_District	UK	54.353329, -2.994200	F
Lhj	Mar_pol_rud	Cambridge	UK	52.188405, 0.117822	M
Liberton	Mar_pol_rud	Edimbourg	UK	55.927015, -3.158666	F
Ncathedral	Mar_pol_rud	Norwich	UK	52.631789, 1.301225	M
RB2	Mar_pol_rud	#N/A	NA	#N/A	M
Rhyl	Mar_pol_rud	Rhyl	UK	53.329780, -3.460486	M
Sevenoaks	Mar_pol_rud	Sevenoaks	UK	51.238861, 0.191621	F
Silverstar	Mar_pol_rud	St_Neots	UK	52.251105, -0.251103	F
Tak1-HK	Mar_pol_rud	#N/A	Japan	#N/A	
Tony	Mar_pol_rud	Selborne	UK	51.091976, -0.944131	F
Warwick	Mar_pol_rud	Warwick	UK	52.202890, -1.628234	F
Waverly	Mar_pol_rud	Glasgow	UK	55.875460, -4.172685	F
Whitley	Mar_pol_rud	Whitley_bay	UK	55.050418, -1.449533	M
Withies	Mar_pol_rud	Guildford	UK	51.210428, -0.691228	F
WT-F	Mar_pol_rud	#N/A	NA	#N/A	F
WT-M	Mar_pol_rud	#N/A	NA	#N/A	

b) Sequencing data on the accessions

The basis of the genomic resource available for *Marchantia polymorpha* is the reference genome constructed on the Tak-1 accession (that I will refer to as TAK1 in this manuscript). The first assembly and annotation were performed by Bowman and collaborators, leading to the reference genome v3.1 (Bowman et al., 2017b), which was then assembled in eight pseudomolecules (reference genome v4, (Diop et al., 2020)). A resequencing and assembly were performed in 2020 (Montgomery et al., 2020) with recent technologies (PacBio long-reads plus Hi-C to get an even better assembly) leading to a fifth version of the genome. The version used for this work was the v6.1 version based on the sequencing of version 5 with the female U-chromosome added (from the Tak-2 accession). This assembly is 222,84 Mbp long and contains 18 335 predicted gene models. All the versions of this genome, gene annotations and other ressources are available on the MarpolBase website (https://marchantia.info/).

TAK1 is an accession from the *ruderalis* subspecies, and reference genome also exist for the two other subspecies (Linde et al., 2020) but are more fragmented: 2 710 contigs for the *montivagans* subspecies reference genome and 2 740 for the *polymorpha* subspecies.

To build the genomic diversity dataset, sampled accessions were all sequenced by Illumina sequencing (2x150 bp paired-end), based on DNA extractions performed on plants grown in axenic cultures in Petri dishes or on potting soil in the green house. These libraries have an expected mean genome coverage of 110X (ranging from 24X to 373X).

To ensure long-term conservation and future distribution of this collection, gemmae were collected from a single individual for each accession and propagated in vitro following sterilization, but some were lost during COVID. Therefore 97 of the sequenced accessions are still available as in vitro culture.

Two accessions from the collection were also sequenced with long-reads technologies. The BoGa-L5 accession from Osnabrück botanical garden was sequenced by Andrea Breautigam's team on a GridION platform and assembled into nine chromosomes anchored on the reference genome (N50=26.28 Mbp; L50=4). The CA accession from California was sequenced by PacBio HiFi in collaboration with the CNRGV platform; and *de novo* assembled into 109 contigs (N50 = 5.99Mbp; L50 = 12). N50 and L50 being statistics representative of the length of a set of contigs: N50 is the length of the contigs for which 50% of the nucleotides of the assembly belong to a contig of the same length or higher, and L50 is the rank of this contig of length N50 (for instance, when contigs are sorted from longer to shorter, you can cover half of the length of the BoGa assembly with 4 contigs, which makes sense since this assembly is composed of 8 chromosome-scale contigs).

Including the reference genome, three long reads genomes are now available for *M. polymorpha* ssp. *ruderalis*, covering a broad geographical range (Japan, USA, Europe). As the synteny plots show (Figure 15) both new genomes cover correctly the *M. polymorpha* reference genome. There is a bit more structural variations between the CA genome and the reference genome, since CA was not assembled by anchoring it on the reference genome.

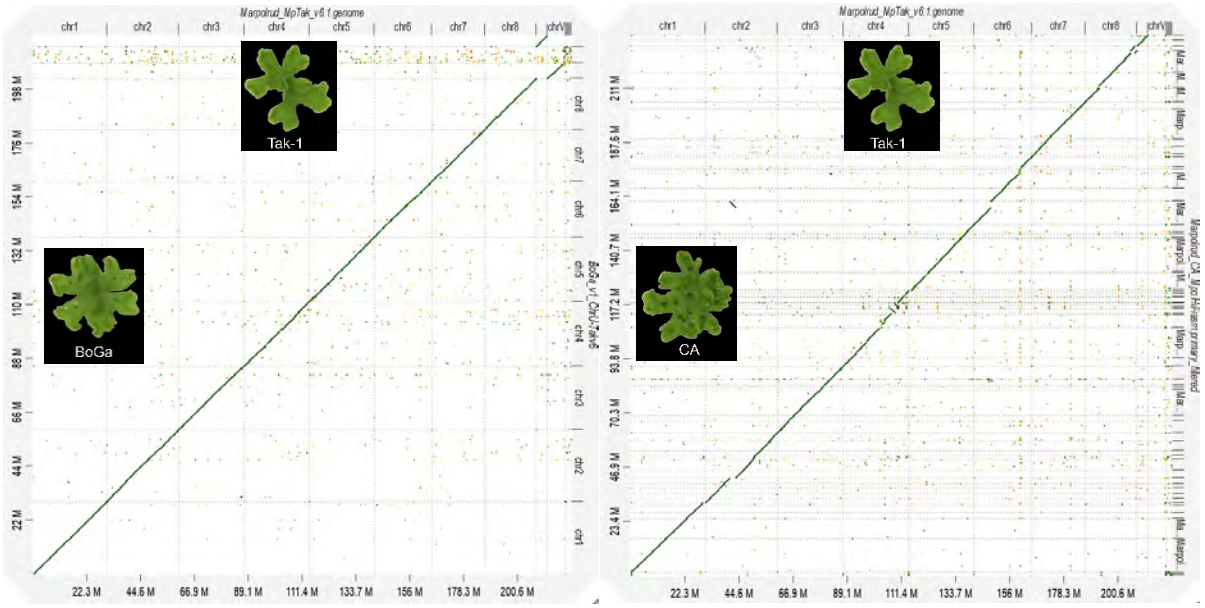


Figure 15: Synteny between the reference genome (TAK1) and the long reads assemblies of BoGa (left) and CA (right). The correspondence were produced with D-GENIES ***(Cabanettes & Klopp, 2018)*** *with default parameters.*

2) Single Nucleotide Polymorphism (SNP) dataset construction

To explore the genomic diversity existing between the accessions, the first approach we used was to perform a SNP calling on all accessions from the three subspecies, based on mapping on the TAK1 reference genome.

a) Mapping on the reference genome

All the sequenced Illumina libraries from the 135 accessions were processed with Trim Galore! v 0.6.5 (Krueger et al., 2021) to remove 3' bases with a quality score lower than 30, trim the Illumina adapters, and only keep the reads longer than 20 bp. Bowtie2 version 2.3.5.1 (Langmead & Salzberg, 2012) was then used to map these reads to the TAK1 reference genome, without permissiveness for discordant and mixed alignments (*i.e.* alignements that do not follow the expected orientation and distance between the two mate pairs, and alignements of one read without its paired read).

The results of the mapping are quite heterogeneous depending on the accession (cf SupData1.1): the mean coverage of the reference genome goes from 0.34X (NILSC10) to 220X (Tou-E) (with an average value over all the accessions of 43.9X, and a median value of 29.8X, while the average value of the expected coverage based on the libraries size is of 110X), while the ratio of mapped reads over the total number of reads in the accessions ranges from 0.4% (NILSC10) to 79% (L2-BoGa and WT-F) (with an average value of 48%, and a median value of 53%). The accessions with poor coverage on the reference genome and a low ratio of mapped

reads are often the same ones and should be taken with caution. We can hypothesize that their libraries might be significantly contaminated with DNA from other organisms, the worst accessions being NILSC10, RES37, RES40, NILSC20 and RES38. Accessions with high coverage are not always the ones with the highest ratio of mapping reads, meaning that they might have large libraries that allow them to cover well the reference genome, but a non-negligible number of their reads do not map and may therefore represent potential contamination. This is the case for Tou-E, Nor-D, Nor-B, or Voe-A. Finally, some accessions like WT-F, L2-BoGa, CAM2 or TAK1 have most of their reads mapping to the reference genome (around 78%) and provide a decent coverage, which means that these libraries mostly contain *M. polymorpha* reads. This is confirmed by the fact that the accessions in this case are mostly the ones cultivated in axenic, in vitro, conditions.

b) SNP calling

Mapping of the 135 accessions on the reference genome were then used to spot polymorphic sites between the reference and each accession. The polymorphic variants with a minimum of 4 supporting reads, a minimum base quality of 30, a minimum variant allele frequency of 0.97, and a p-value (of variant read count vs. expected baseline error) threshold of 0.01 were then called with VarScan.v2.4.2 (Koboldt et al., 2009, 2012). The resulting VCF combines polymorphic positions for all 135 accessions, leading to a total of 13 024 932 indels and SNP sites. All accessions were screened for the number of missing data at these sites. RES37 and NILSC10 had 99% of missing data over all the sites and were therefore filtered out from all subsequent analyses. Other accessions with poor mapping rates also had a lot of missing data (73% for RES40, 66% for WT-M, 64% for NILSC17) but the values were not as dramatic as the ones of RES37 and NILSC10 and they were therefore kept in the dataset.

The VCF file of all *M. polymorpha* accessions was therefore filtered to discard the 2 accessions of very low quality and the indels, and to keep only biallelic sites with information in at least 50% of the accessions, leading to a total of 12 519 663 SNPs in the 133 remaining accessions. There is therefore a density of 1 SNP every 18 bp in this dataset that takes into account the differences between the three subspecies and the intra-subspecies diversity. Independent SNP files were also made for each of the subspecies, with the same filters (no indels, only biallelic

sites and maximum 50% of accessions with missing data at one site). The 13 ssp. *polymorpha* accessions had 1 883 260 SNPs, the 16 ssp. *montivagans* accessions 754 645 and the 104 ssp. *ruderalis* accessions 5 414 844 SNPs, with a SNP density of approximately 1 SNP every 40 bp. By comparison, the *A. thaliana* SNP dataset of the 1001 genomes consortium **(Alonso-Blanco et al., 2016)** has a density of approximately 1 SNP every 10 bp, and the *M. truncatula* SNP dataset of 285 accessions on the version 5 of the reference genome **(Epstein et al., 2022)** has a density of approximately 1 SNP every 8 bp. Most of the population and quantitative genetic analyses will therefore focus on *M. polymorpha* ssp. *ruderalis* given the largest sample size in this subspecies.

c) Genome-wide patterns of linkage disequilibrium

Based on SNPs from the subspecies *ruderalis*, the linkage disequilibrium (LD) decay was determined using PopLDdecay **(C. Zhang et al., 2019)** with a MAF of 0.05 and a maximum pairwise SNP distance of 40 kb, on the whole genome and on each chromosome separately (Figure 16). The one-half linkage disequilibrium (LD) decay ranged from 2.5 (chromosome 8) to 4.7 (chromosome 1) kb across the eight *M. polymorpha* ssp. *ruderalis* autosomes, with a genome-wide average of 3.6 kbp. To give a perspective the dense SNP dataset of A. thaliana and M. truncatula give half LD decay distances of respectively 0.9 kb and 0.6 kb (Appendix A). The linkage disequilibrium is therefore not very high in *M. polymorpha*, suggesting an equilibrium between clonal and sexual reproduction in this liverwort. This conclusion was also drawn in a recent population genomics study, on a local sampling of 23 *M. polymorpha* accessions from Ontario in Canada **(Sandler et al., 2023)**.

This relatively low LD means that genetic mapping can be implemented through genome-wide association approaches, at a fine scale.

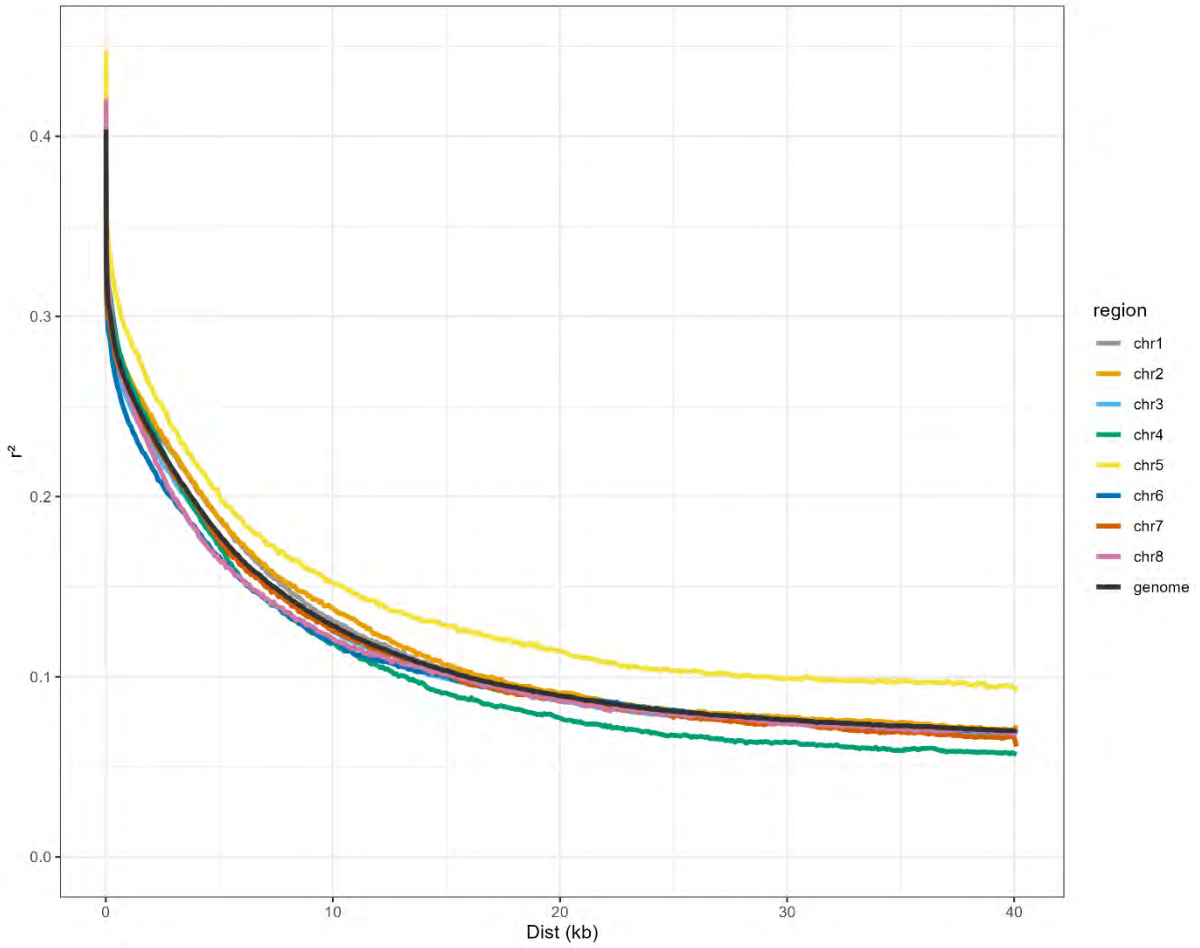


Figure 16: Linkage disequilibrium decay in Marchantia polymorpha, for each chromosome, and for the whole genome (black line).

3) Study of the genetic structure of the *M. polymorpha* collection

The genetic relationship between accessions was inferred using SNP-based analyses, in order to verify the correct assignation of each accession to its subspecies and to see whether there is a link between geographic distances and the genetic relationships between accessions. The first approach was a SNP-based phylogeny on biallelic and multiallelic SNPs shared by the 133 accessions with a maximum of 5 accessions with missing data per site. These sites were then pruned with the R package SNPRelate (X. Zheng et al., 2012) for a linkage disequilibrium (LD) < 0.3 on 1 500bp windows, leading to a reduced list of 107 934 representative and non-redundant SNPs. The SNP based phylogenetic tree was inferred with IQ-TREE version 2.1.2 (Minh et al., 2020) using the SYM+ASC+R6 model (chosen by ModelFinder (Kalyaanamoorthy et al., 2017)) with SH-like approximate likelihood ratio test and ultrafast bootstrap (with 10 000 replicates). The tree was rooted on the *montivagans* subspecies, following the topology suggested by Linde et al **(Linde et al., 2020)**, and by our Orthofinder analysis on genes from all accessions (cf

chapter 3 and Appendix F). The three subspecies are well separated on the tree (Figure 18), and all the accessions were correctly assigned. Only Pra-B, that belongs to the *polymorpha* subspecies, displays some similarity features from the *ruderalis* subspecies as well, suggesting an hybridation. This is quite intriguing since the only description of a crossing experiment leading to viable spores was between the *polymorpha* and *montivagans* subspecies **(Linde et al., 2020)**. The branches of the tree were colored depending on the geographic area of origin of the accessions (Figure 18), showing that there is no clear geographical structuration of the population. This was verified by a Mantel test on the genetic distance matrix (based on allele counts calculated with PLINK2 **(Purcell et al., 2007)**) and the geographical distance matrix (geodesic, calculated with *geosphere* R package) between 96 georeferenced accessions (Mantel statistic r = 0.087, p = 0.074, with *vegan* R package) (Figure 17). This weak correlation between geographic and genetic distances, could be due to the human-caused spread of Marchantia, notably through horticultural trade. Indeed, *M. polymorpha* abounds in greenhouses or in pots of ornamental plants **(Marble et al., 2017)**, and might be transported around the world by this means.

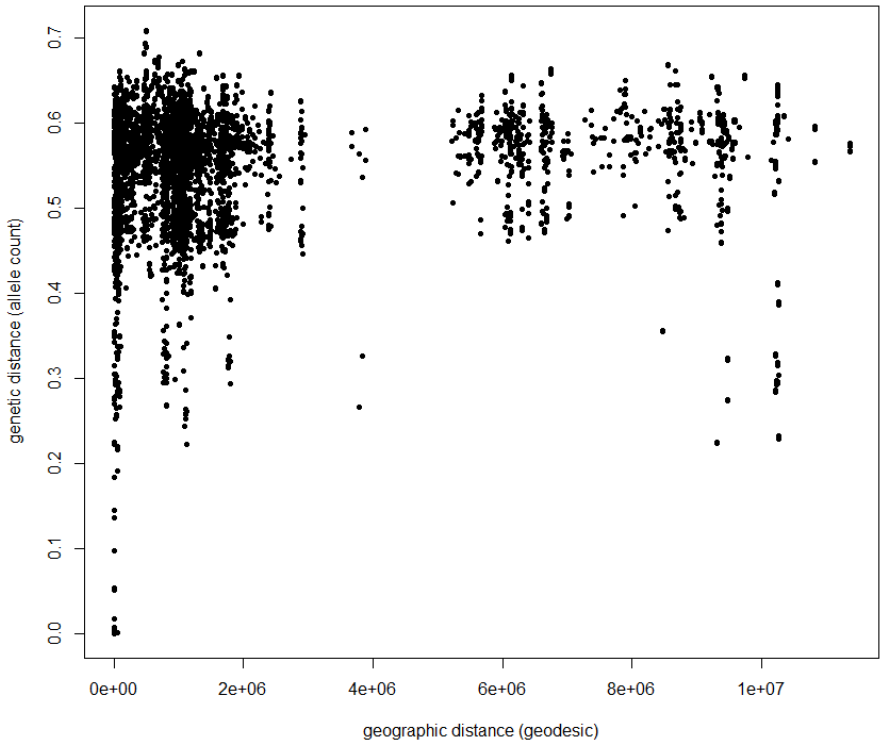


Figure 17: Geographic distance between accessions vs genetic distance between accessions. The genetic distance is represented by the number of sites with different alleles in two accessions. Geographical distance and genetic distance are not significantly correlated in our collection.

The population structure within the ruderalis subspecies was also investigated, using the fastStructure software (Raj et al., 2014). FastStructure uses a model-based approach to infer population structure from large SNP dataset, similarly to STRUCTURE or ADMIXTURE. The analysis was performed on 95 non redundant accessions from the *M. polymorpha* ssp. *ruderalis*, meaning that for groups of accessions sampled close to each other and with high genetic similarity only one was kept (for instance Tou-E for the Tou group). When this redundancy was not taken into account, fastSTRUCTURE was always assigning the clusters to these small groups of accessions and was not describing the higher level of structure. The SNPs with a missing frequency < 20% and a minor allele count of 6 accessions (that correspond to a MAF >6% when there is no missing data at the site) were selected and pruned with PLINK2 on 10 kb windows with a step size of 1 and a LD threshold of 0.5. The 458 346 resulting SNPs were used for analysis with fastStructure v1.0. Population numbers (K) from 2 to 7 were tested with 5 cross-validation test sets, and the chooseK algorithm pointed to an optimal number of clusters between 2 and 4. The absence of sharp genetic structure already observed, being most likely the result of recurrent gene flow, is probably the reason why the software does not provide a clear-cut definition of the number of genetic groups in *M. polymorpha* ssp. *ruderalis*. I therefore chose to represent the assignation of the accessions from the *ruderalis* subspecies to 3 groups (Figure 18). The distribution of the accessions in the groups is quite consistent with their distribution on the tree. Some accessions like Tou-E or Ber-D seem to be admixed between different groups, suggesting gene flow across the broad geographical range of *M. polymorpha* ssp. *ruderalis*. We also observe clusters of high genetic similarity in our population, but at very local scales (with for instance the Tou or the Ber populations).

Similar conclusions were drawn with a recent study on a local sampling of 23 *M. polymorpha* accessions from Ontario in Canada (Sandler et al., 2023), that found close relatedness between geographically distant accessions and signs of frequent gene flows.

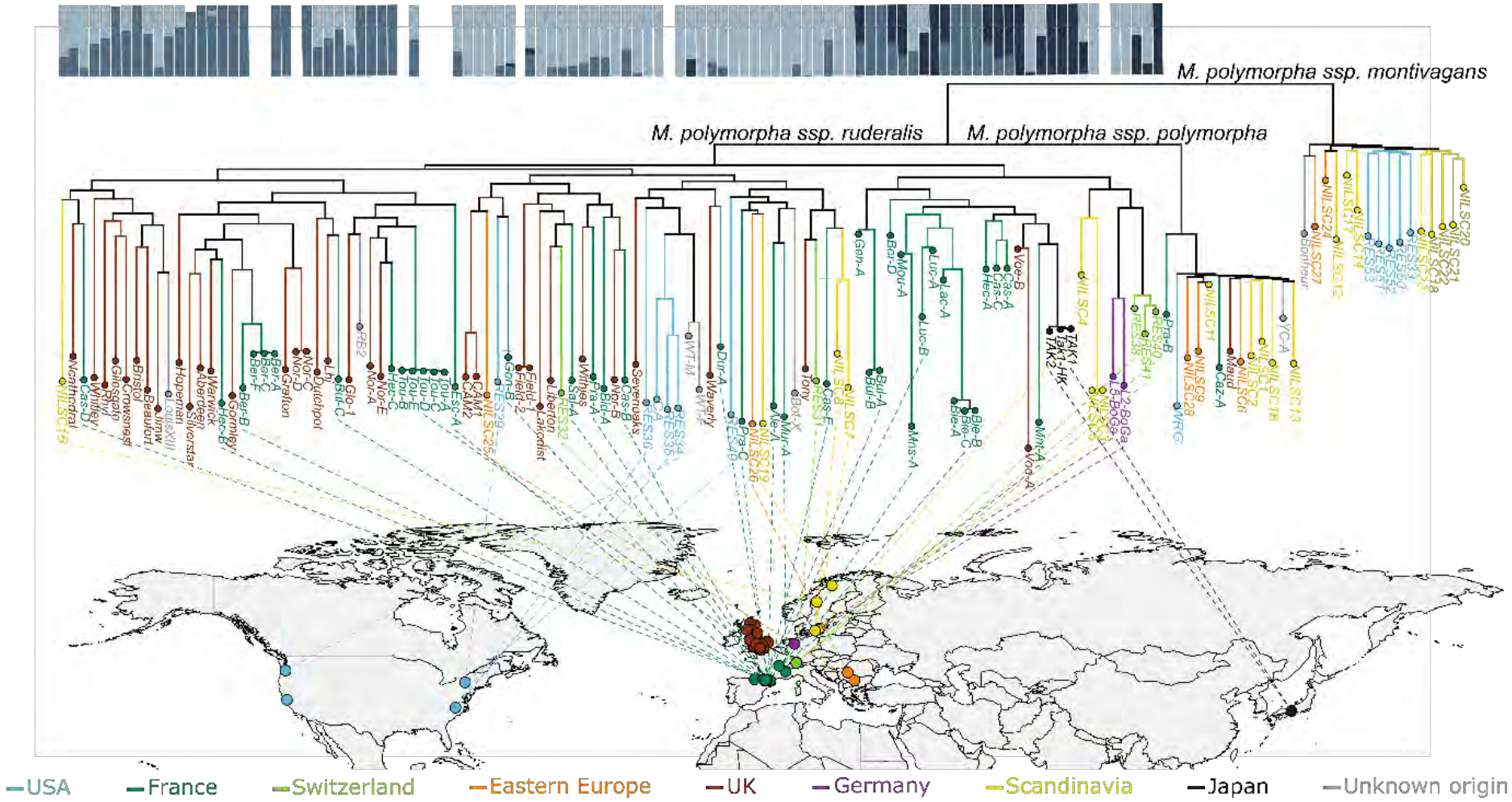


Figure 18: Phylogenetic tree of the accessions from the M. polymorpha collection, with their geographical origin. For the accessions from the ruderalis subspecies, the assignation to three different genetic groups by a FastStructure analysis is represented on the top of the tree.

II) Patterns of selective pressures on *M. polymorpha* ssp. *ruderalis*' genes

1) SNP distribution and allele frequency spectrum in *M. polymorpha*

The *M. polymorpha* ssp. *ruderalis* SNPs were examined according to their predicted functional effect in order to understand the selective pressure on the different types of variants. To do so, the SNPEff software **(Cingolani et al., 2012)** was used on the 5 414 844 sites called in the *ruderalis* subspecies to predict the potential effect of each SNP, using the TAK1 v6 gene annotation. We focused on variants located in 3' UTR, 5' UTR, intergenic and intronic regions and on variants with a specific effect. These effects can be missense (producing a different amino acid) or synonymous (producing the same amino acid) mutations, 5' UTR premature start codon gain (variant in 5' UTR region producing a three base sequence that can be a START codon), splice acceptor or donor (affect mRNA splicing), start lost, stop gained, stop lost or stop retained (change of nucleotide leads to another stop codon).

Among the main SNPs categories, 80% were in intergenic regions. In genic regions, SNPs were similarly frequent in intronic (9%) and in non-intronic (11%) regions – that is exons and UTR regions – (Figure 19). The genome-wide allele frequency spectrum based on these SNP categories' minor allele frequency is clearly L-shaped, as expected under neutral evolution. We observed an excess of rare missense, 5' UTR, 3' UTR and the gain of premature start codon in 5' UTR, relative to SNPs located in intergenic regions, introns, or synonymous SNPs. This distribution most probably reflects an excess of purifying selection on the formers (Figure 19). Loss-of-function SNPs were predominantly leading to the gain (51%) and less frequently to the loss (25%) of stop codons. The genome-wide allele frequency spectrum of such SNPs was also L-shaped, with an excess of rare stop gains and splice acceptor SNPs (Figure 19). This reflects an excess of purifying selection against variants leading to molecular phenotypes associated with truncated proteins and retained introns.

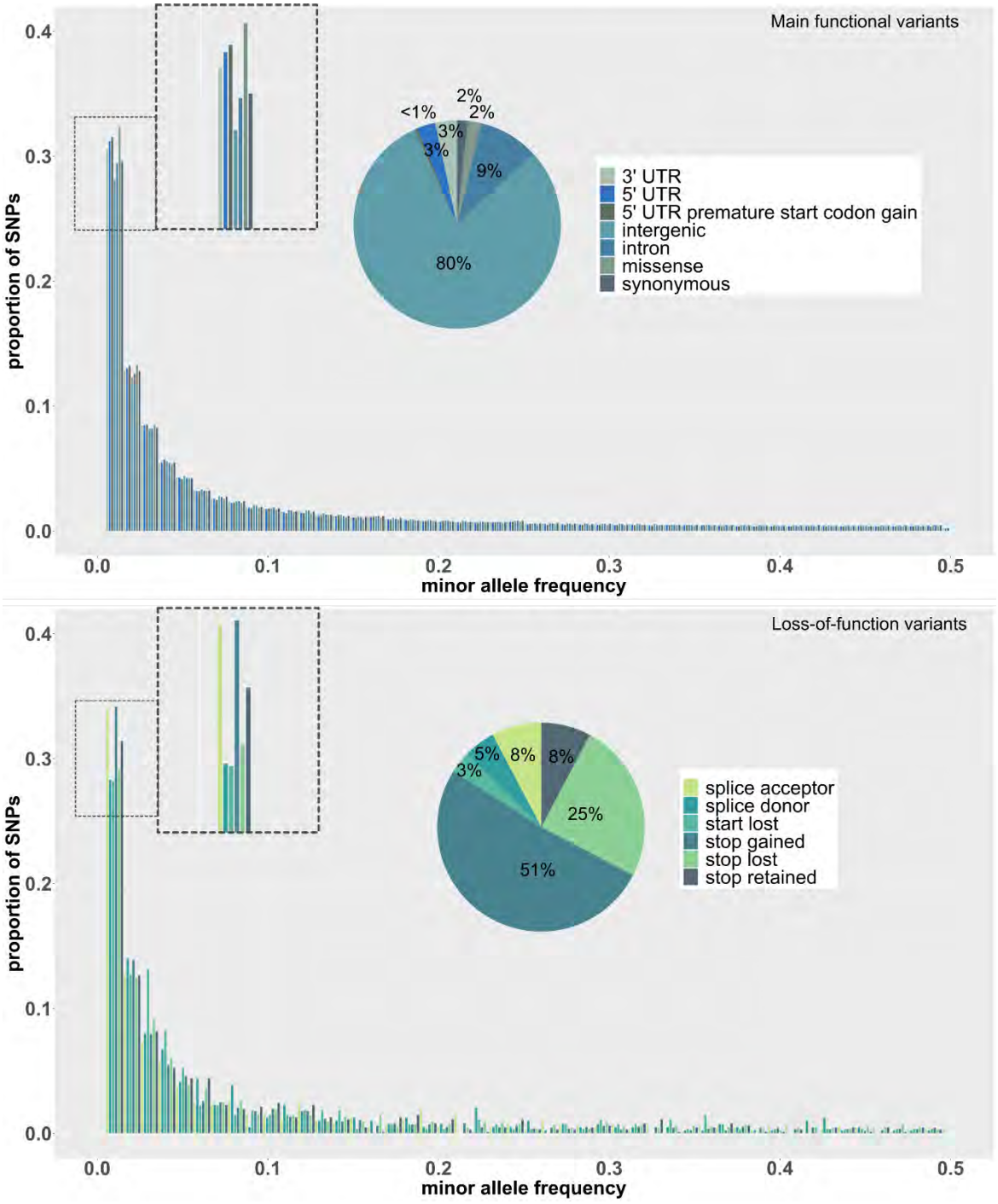


Figure 19: Allele frequency spectrum for different categories of SNP, based on minor allele frequency. On the upper panel, 5 439 674 main functional variants (4 363 099 intergenic, 509 964 in introns, 115 218 missense, 88 163 synonymous, 183 858 3'_UTR, 154 821 5'_UTR and 24 551 5'_UTR_premature_start_codon_gain), were used to construct a site frequency spectrum. On the lower panel, 7 823 loss of function variant (4 010 stop gained, 1 948 stop lost, 608 stop retained, 606 splice acceptor, 385 splice donor and 266 start lost) were used to construct another site frequency spectrum based on rare functional variants.

2) Presentation of the neutrality test statistics used

a) General concept behind Tajima's *D*, Fay and Wu's *H* and Zeng's *E*

As mentioned in the introduction and as it can be seen in the first part of this chapter, selection processes leave a pattern on the polymorphic sites by modifying the allele frequencies in the population at the site under selection and on linked sites. Indeed, under the neutral model the distribution of the mutation frequencies (or site frequency spectrum: SFS) is L shaped as shown in Figure 20 **(Crow & Kimura, 2009; Wakeley, 2009)**. When a site is under balancing selection (and so are linked sites), this distribution will shift towards mutations with intermediate frequencies because both alleles are "equally" selected. For a selective sweep, the SFS will have an excess of high frequency mutations (the selected alleles and the ones hitchhiking with it) plus an excess of low frequency mutations (the mutations that appeared independently in individuals after the spread of the selective sweep in the population). Finally, in the case of background selection, the mutations will be strongly counter selected, which will lead to an excess of rare mutations (as shown in Figure 20 fourth panel and Figure 19 for missense and stop gained variants for instance).

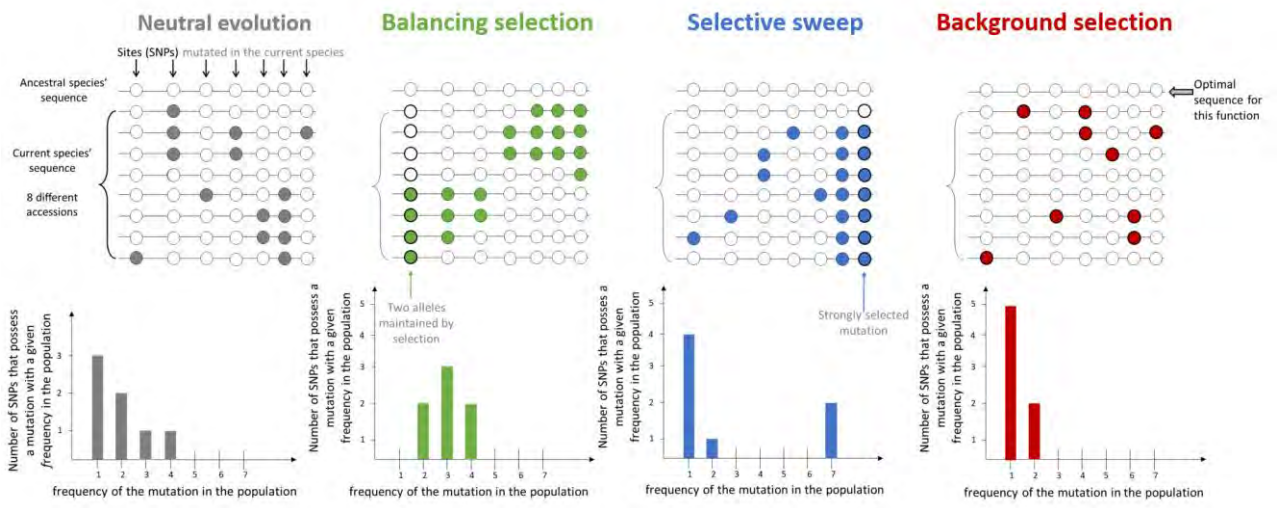


Figure 20: Detail of the different type of selective forces and their effect on the frequency spectrum in a genomic region. The ancestral alleles (white) and mutations (color) are represented for 7 SNPs, in 8 haplotypes. Depending on the type of selection effective, the mutations will segregate at low, intermediate or high frequencies, and the SFS will change.

Some statistical tools were developed by population geneticists to screen genomes for the SFS left by these selection pattern, by using summary statistics derived from the population

genetics parameter θ=4Nμ. We have used three of them in our studies: Tajima's *D*, Fay and Wu's *H* and Zeng's *E*. They are calculated as follows:

$$D = \frac{\theta_\pi - \theta_s}{\sqrt{Var(\theta_\pi - \theta_s)}}$$

$$H = \frac{\theta_\pi - \theta_L}{\sqrt{Var(\theta_\pi - \theta_L)}}$$

$$E = \frac{\theta_L - \theta_s}{\sqrt{Var(\theta_L - \theta_s)}}$$

The general principle of these statistics is to evaluate the difference between two θ estimators that allows to detect deviations from the neutral model, where gene polymorphisms θ theoretically only depends on the effective population size N and on the mutation rate μ (θ=4Nμ). The three different θ estimators, θs, θπ and θ_L all reflect the polymorphism in the population but are sensitive to changes in different parts of the frequency spectrum either due to changes in the selection regime or in the demography. θ_s accounts for changes in the number of rare variants in the population, θπ for changes in the number of variants with intermediate frequencies and θL for changes in the number of derived variants with high frequencies. Since the different types of selection act differently on each part of the frequency spectrum, contrasting the θ estimators two by two with the D, H and E statistics allows determining which selective force(s) acted on a given genomic region.

Tajima's *D* (Tajima, 1989) compares the number of pairwise differences between individuals (θπ) to the total number of polymorphic sites in the population (θs), in a given genomic region. Rare alleles participate less in the pairwise differences between individuals because they are present in less individuals. Therefore, a negative Tajima's *D* represents an excess in rare alleles and therefore a zone under strong directional or purifying selection. On the contrary, a positive *D* means that there is a lot of pairwise polymorphisms between individuals, and therefore an excess of intermediate frequency alleles, corresponding to balancing selection (Figure 20).

Fay and Wu's *H* (Fay & Wu, 2000) compares the number of pairwise differences (θπ) to the mean number of mutations accumulated since the last common ancestor of the individuals (θ_L). If *H* is negative, it means that there is an excess of high frequency derived alleles in the

population, which is an indication of selective sweep (the selection of a new allele). This statistic is quite powerful to detect on-going selective sweeps, but not fixed selective sweeps. A positive value of *H* represents the maintenance of different intermediate frequency alleles, and therefore a balancing selection.

Zeng's *E* (Zeng et al., 2006) also uses the mean number of mutations accumulated since the last common ancestor (θ_L), but it compares it to the total number of polymorphic sites in the population (θs). A negative value of *E* represents an excess of low frequency derived alleles, which is a signature of negative selection (i.e. background selection on sites linked to deleterious mutations which are under purifying selection). In this type of selection most new mutations are deleterious and counter-selected, leading to a deficit in derived alleles. A positive *E* represents an excess of high frequency derived alleles in the population, due to a strong selection of this type of alleles. Selective sweeps can therefore be detected, but in this case the statistics has more power to detect already fixed selective sweeps.

It is therefore possible to differentiate between selective sweep, background selection and balancing selection by combining these three statistics. Tajima's *D* indicates if the selection is balancing (positive *D*) or directional (negative *D*), and then the *H* and *E* allow discriminating between a positive selection (selective sweep: positive *E*, negative *H*), or a counter-selection (background selection: negative *E*) of derived alleles.

Nevertheless, the allelic frequencies in a population are also impacted by demographic events, like bottlenecks (a small part of the initial population leads to the studied population, which lead to a deficit in rare alleles) or a population expansion (the population goes from a small group of individuals and expands rapidly, therefore most sequences are identical, to the exception of rare mutations that appeared after). The influence of these demographic events on the statistics is specified on the Figure 21. A way to compensate for this bias is to consider the statistics value obtained for a given genomic region relatively to other values obtained in all the other regions of the genome (in the case of this study, since we will focus on genes, the genome-wide gene-based distributions of *D*, *H* and *E* will be considered). Demographic forces affect the frequency spectrum the same way in the whole genome, whereas selective forces affect specific regions. Even if the genome-wide distribution of the statistics is not centred on zero for all the genomic regions, due to a genome-wide (i.e. for all the studied genomic regions) excess (population expansion) or deficit (population bottleneck or structure) of rare alleles,

genomic regions under selection will still tend to show more extreme statistics values. As shown in Figure 21 (bottom panel), in *M. polymorpha* ssp. *ruderalis* the average value of gene-based Tajima's *D* and Zheng's *E* statistics were negative (*D* = -0.356, *E* = -0.317), indicating a genome-wide excess of low-frequency variants. This signature suggests demographic expansion (perhaps accelerated by recent horticultural trade) with low population substructure, in agreement with the phylogenetic and population structure analyses.

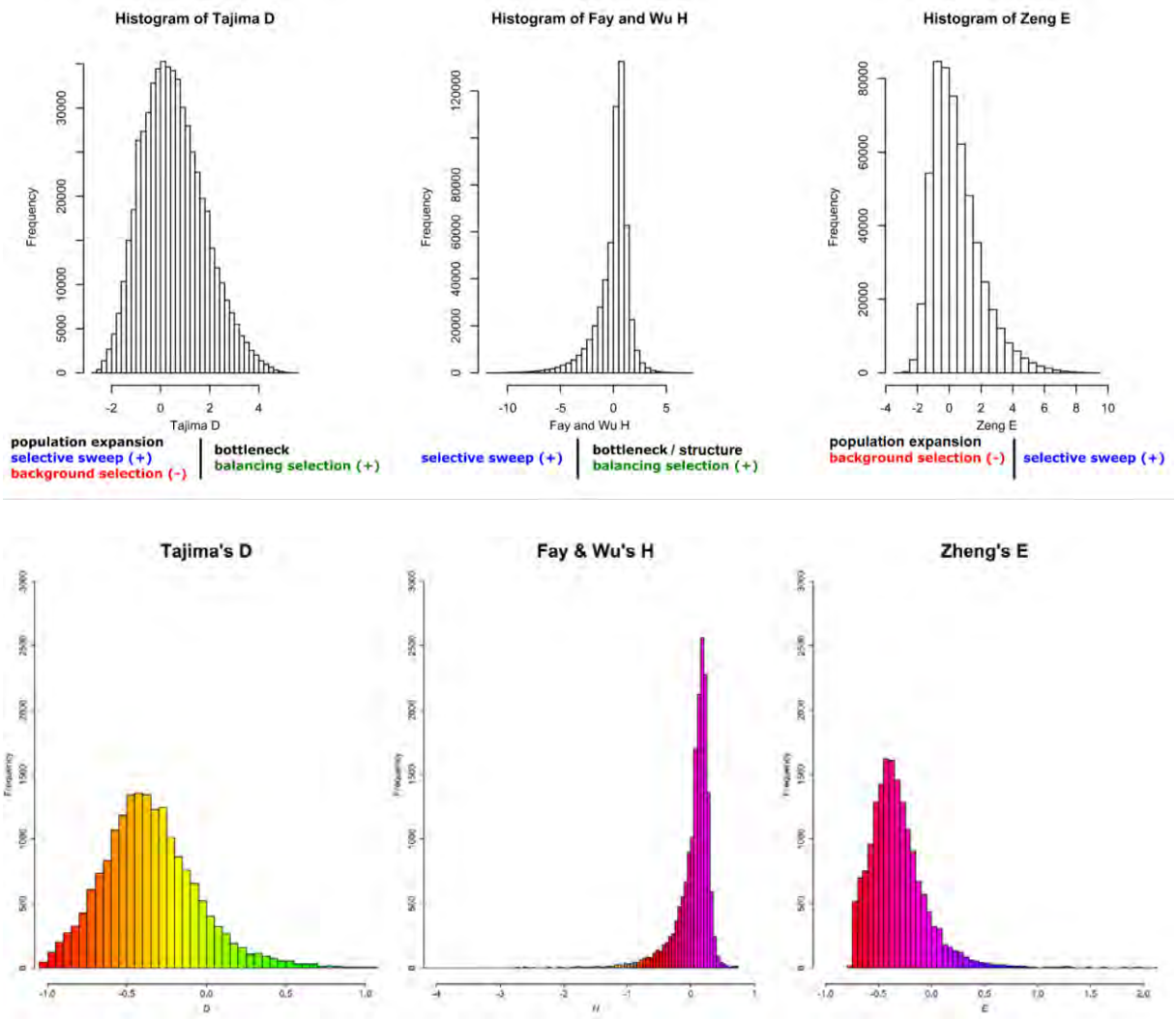


Figure 21: Empirical distributions of Tajima's D, Fay and Wu's H and Zeng's E (top, Source: Léa Boyrie's PhD manuscript) and real distribution in M. polymorpha ssp. ruderalis based on 18 140 genes for D and 17 027 genes for H and E (bottom). For the empirical distributions, the selective forces corresponding to each side of the statistics distribution (negative or positive values) are shown under the distributions, as well as the demographic forces that can also impact the statistics.

b) Considerations on methods for calculating the *D*, *H* and *E* statistics

Usually, *D*, *H* and *E* statistics are calculated based on genomic windows that contain several SNPs. However, if the statistics has been calculated on a given genomic window (a gene for

instance), to know its value in a subset of the genomic window (the gene's exons for instance), the entire calculation must be redone. Moreover, the calculation of these statistics does not take into account the presence of missing data and should therefore be used only on sites without any missing data for any accession. This greatly reduces the number of SNPs that can be considered for the calculation. To circumvent these limits, we modified a DHE calculation script written based on the variance estimators detailed in Zeng et al. 2006 **(Zeng et al., 2006)**, which takes as input a SNP matrix recoded in 0 and 1 (representing ancestral and derived alleles or major and minor alleles, in a haploid organism). The window-based calculation of the DHE in this script was changed to a SNP-based calculation, on which we could then implement a correction for the true sample size at each site, inspired by the work of Ferretti et al. **(Ferretti et al., 2012)**. The values of the statistics over a given genomic window could then be calculated by averaging the *D*, *H* and *E* values obtained on SNPs over the window. Calculating these indicators on different windows would then only require a simple averaging, instead of a complex calculation of the statistics from scratch.

The basis of the *D*, *H* and *E* statistics calculation, that is the previously mentioned θ estimators, are usually calculated as follows:

$$\theta_s = S \Big/ \sum_{i=1}^{n-1} 1/i$$

$$\theta_\pi = \binom{n}{2}^{-1} \sum_{i=1}^{n-1} i(n-i) \times S_i \; = \; \binom{n}{2}^{-1} \sum_{i<j}^{n-1} d_{ij}$$

$$\theta_L = \frac{1}{n-1} \sum_{i=1}^{n-1} i \; \times S_i$$

With n the number of "haploid individuals" (n is two times the number of individuals in a diploid species), S_i the number of sites comprising i derived alleles at the locus (the frequency spectrum), S the total number of polymorphic sites at the locus (i.e. the number of SNPs) and d_{ij} the number of differences between two sequences (i and j) at the locus.

Thus, θ_s represents the number of polymorphic sites at the locus, corrected by the number of sequences studied (*i.e.* the average total branch length underlying the sampled sequences), θ_π represents the mean number of pairwise differences between two individuals at the locus and

θ_L represents the mean number of mutations accumulated since the last common ancestor of the sequences (this indicator requires to know the derived or ancestral state of the alleles).

When calculated on only one SNP and not on an entire window, these estimators can be simplified as:

$$\theta_s = 1 \Big/ \sum_{i=1}^{n-1} 1/i$$

(S replaced by one since there is only one segregating site)

$$\theta_\pi = \frac{number_{ancestral\ alleles} \times number_{derived\ alleles}}{n \times (n-1)/2}$$

n×(n-1)/2 accounts for the number of possible pairwise combinations between individuals

$$\theta_L = number_{derived\ alleles\ at\ this\ site}/(n-1)$$

The number of derived alleles is easy to calculate by summing the recoded values of the SNPs (since derived alleles are coded as 1 and ancestral alleles as 0). The number of ancestral alleles can be deduced by subtracting the total number of alleles by the number of derived alleles.

The correction for missing data is implemented by a correction of the n (number of haploid individuals) at each site:

$$n = number_{individuals} - number_{\ individuals\ with\ missing\ data\ at\ the\ site}$$

The scripts gathering all these calculations are available at the following link: https://figshare.com/s/715a36bc7585d46d0279 folder DHE_calculation_scripts.

A simulation of 6000 loci (with the ms software **(Hudson, 2002)** assuming various theta, rho (recombination), alpha (demographic expansion parameter) and nsite values) was used to verify that the averaging of *D*, *H* and *E* values over SNPs was equivalent to the window-based calculation. The statistics were calculated on all simulated loci with the two approaches and plotted, leading to correlations of 1, 1 and 0.99 for *D*, *H* and *E* respectively between the window calculation and the SNP calculation.

The efficiency of the missing data correction was also verified, by artificially adding 50% of missing data in the simulated loci. The *D*, *H* and *E* statistics calculated per site and averaged on these loci with missing data were then compared to the same statistics calculated by the

window-based approach on the same loci without missing data. This comparison showed that the correction implemented allowed to approach the real values of the statistics despite the important rate of missing data (Figure 22).

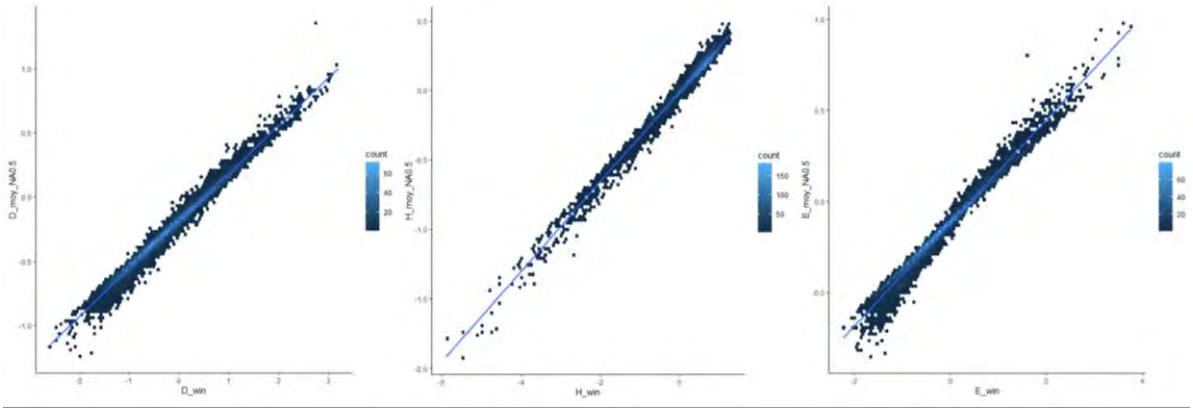


Figure 22: comparison of the D, H and E statistics calculated on the simulated loci by the window-based approach (x axis) and on the same loci with 50% NA, with the averaging-based DHE calculation that implements a missing data correction (y axis). The correlation coefficients between the two approaches are of 0.99, 0.99 and 0.99 for D, H and E respectively.

This modification of the calculation method for the *D*, *H* and *E* statistics is accurate and can be used on our data. It provides two main advantages:

- a flexibility of calculation on different windows: once the statistics are calculated on each site (SNP) of the genome, a simple averaging of the values can be done easily to get the *D*, *H* and *E* values on different genomic regions.

- an accuracy of calculation even for the sites with missing data (by taking into account the real sample size of individuals for each SNP), allowing to use more SNP, since the ones with incomplete information for some accessions do not have to be discarded anymore.

3) Calculation of *D*, *H* and *E* statistics in three plant species

One of the objectives of my PhD project was to study *M. polymorpha* on two scales: the microevolutionary one (by comparing the different accessions) and the macroevolutionary one (by comparing different species). To make a first step towards this goal we decided to explore the selective forces acting on genes in three species: *M. polymorpha* and the two angiosperm model species *A. thaliana* and *M. truncatula*. These two species do not cover optimally the diversity existing in angiosperms (they both are from the rosid clade), but large SNP datasets are available for them, making them good candidates for a first trial of interspecific comparison

of intraspecific gene selection signatures. For the three species, we processed following the same rationale to calculate the *D*, *H* and *E* statistics.

a) *D*, *H* and *E* calculation on *M. polymorpha* ssp. *ruderalis* SNPs

To calculate *H* and *E*, the knowledge of the ancestral and derived allele for each SNP is required (the unfolded SNP dataset). To determine the unfolded SNP dataset in our collection of 104 *M. polymorpha* ssp. *ruderalis*, the two other subspecies from the *M. polymorpha* complex were used. The polymorphic positions from the *ruderalis* subspecies were examined in the two other subspecies and the sites fixed with the same allele in all the 13 accessions from ssp. *polymorpha* and all the 16 accessions from ssp. *montivagans* (with a tolerance for missing information in one accession from each outgroup subspecies) allowed to determine the ancestral allele of the *ruderalis* subspecies. Indeed, if the two other subspecies bear the same allele at one site, the parsimonious hypothesis is that this allele is the ancestral one, and that the other allele that is present at this site in the *ruderalis* subspecies is the derived one, which appeared after the divergence of the *ruderalis* subspecies. This ancestral-derived state could only be determined in 1 344 013 polymorphic sites from the *ruderalis* subspecies, that were then recoded as 0 for ancestral allele and 1 for derived allele and used to calculate the *D*, *H* and *E* statistics (this will be referred to as the unfolded SNP dataset). For the calculation of Tajima's *D*, the folded SNP dataset (5 414 844 SNPs) was used as well, because there is no need to know the ancestral or derived state of the SNPs. After calculation of the statistics with the custom R script, they were averaged over the 18 920 gene models present in the TAK1 reference genome.

The two values of D (with the folded SNP dataset and with the reduced unfolded SNP dataset) can be compared when considering the selection signature in a genomic region, to see if the selection of SNPs with an inferable ancestral allele creates a bias.

b) Statistics calculation in *A. thaliana* and *M. truncatula*

For *Arabidopsis thaliana*, the SNP resources developed by the 1001 genomes project **(Alonso-Blanco et al., 2016)** was used. This dataset contains 11 458 975 SNPs coming from 1135 *A. thaliana* accessions mapped on the TAIR10 genome. To be able to infer the ancestral allele of a subset of these SNP, sequencing data from 30 accessions from *A. halleri* and 29 accessions from *A. lyrata* were downloaded from the NCBI's SRA (respectively PRJNA592307 and

PRJNA357372) and mapped on the TAIR10 reference genome. To do so the reads were processed with Trim Galore! similarly to what was done for the reads of our *M. polymorpha* collection, and then mapped with Bowtie2 version 2.3.5.1 with permissiveness for discordant and mixed alignment, and with an alignment score lower than the minimal score threshold -1.5 + -1.5*L, to allow mapping of reads from distant species. Coverage of the TAIR10 genome with these mapping was decent, with an average of 27X (ranging from 8X to 64X). The polymorphic variants present in *A. thaliana* were called with VarScan.v2.4.2 in the two other subspecies, with a minimum of 4 supporting reads, a minimum base quality of 30, a minimum variant allele frequency of 0.97, and a p-value threshold of 0.01 when comparing the significance of the variant read count to the baseline error, leading to 9 500 949 sites found in the *A. halleri* and 9 206 703 in the *A. lyrata*. The sites with a maximum of 30% of missing data in each of the outgroup species and with the same allele fixed in all the accessions were the ones forming the unfolded SNP dataset in *A. thaliana*. This led to 3 641 556 SNPs with ancestral allele information, on which the *D*, *H* and *E* were calculated. Tajima's *D* was also calculated on the whole dataset of 11 458 975 SNPs. The statistics were averaged over 26 530 *A. thaliana* gene models.

For *M. truncatula* the SNP dataset based on mapping of 317 accessions (among which 285 are from the *M. truncatula* species, the other ones being from 23 other Medicago species) **(Epstein et al., 2022)** on the version 5 of the A17 genome **(Pecrix et al., 2018)** was used. Eight accessions belonging to the closely related species *turbinata, italica, doliata, littoralis, turbinata, tricycla and soleirolii* were taken as outgroups. The ancestral allele was determined for sites that were fixed the same way in all 8 outgroup accessions, with a tolerance for missing data in one accession only, leading to a dataset of 12 129 573 unfolded *M. truncatula* SNPs, on which *D*, *H* and *E* were calculated. Tajima's *D* was also calculated on the whole dataset of 50 885 956 biallelic SNP with a minimum base quality of 30, no half calls and a maximum 50% of missing data. Contrary to *M. polymorpha* that is haploid and *A. thaliana* that is a diploid selfing, *M. truncatula* which has a high selfing rate in local population **(Siol et al., 2008)** showed higher levels of heterozygosity (panmixia was rejected by a chi-square test for approximately 50% of the SNPs). The script used to calculate the statistics was therefore adapted by changing the n parameter (the number of haploid individuals, which is two times the number of diploid individuals). The statistics were averaged over 49 538 *M. truncatula* genes models.

Since we aim at comparing the three species, we focused here on the genes and the forces of selection acting on them, but the calculation of the *D*, *H* and *E* statistics on all SNPs from these three species can be a resource to study precise regions of interest.

4) Evaluation of intraspecific selective pressures in *M. polymorpha* genes

a) SNP distribution in the light of the unfolded SNP dataset and diversity indicators

Since the ancestral and derived alleles are predicted for all sites for which it was possible, we can now assess the site frequency spectrum based on the frequency of the derived allele. The type of variants with an excess of rare frequencies of mutated allele are mostly the same as in the SFS based on the minor allele frequency, being variants in the UTR regions, missense variants and stop gained variants (Figure 23). This SFS also shows that start lost and stop gained variants that appeared in the *ruderalis* ssp. are strongly counter selected (they have an excess of rare variants).

Regarding derived alleles that almost got fixed in the population (high frequencies), most categories of main functional variant are all fixed in a similar manner. There is a non-negligible proportion of missense variants fixed, representing changes in the amino acids that do not affect too badly the structure of the protein, or could even represent a beneficial change of amino acid sequence. Among proteins that contain a selected missense variant (frequency over 90%), there are multiple pentatricopeptide repeat proteins (MpPPR_26, MpPPR_16, MpPPR_3…), multiple cupins (MpCupin2, MpCupin14, MpCupin26…), multiple peroxidases (MpPOD15, MpPOD58, MpPOD97…), or NLRs (MpNBS-LRR4, MpNBS-LRR11) (SupData1.2). For the loss of function variants, stop lost and stop retained have been more positively selected than other types. There is a total of 84 genes containing LoF variants positively selected in the population (frequency over 90%). Among the genes with stop lost variants are a cupin (MpCupin 58), an ethylene responsive transcription factor (MpAP2B3-2) or a peroxidase (MpPOD114). Genes with stop gain are for instance a LURP1 related protein, two histine kinases (MpPAS6 and MpCache3), or an autophagy protein (MpATG5) (SupData1.3, sheet 1). There is also loss of function variants under balancing selection in the population, in a total of 538 genes (SupData1.3, sheet 2). Among the genes with these balancing LoF variants that are globally under strong global balancing selection, are the NBS-LRR11 (with two stops lost in 24% of the population), some peroxidases: MpPOD118 (stop lost in 63% of the population) and MpPOD1 (stop lost in 65% of the population), or a chitinase MpGH19.11 (with a stop lost variant

in 36% of the population and a stop gained variant in 25% of the population). Many of these variable genes have multiple LoF variants affecting their sequence.

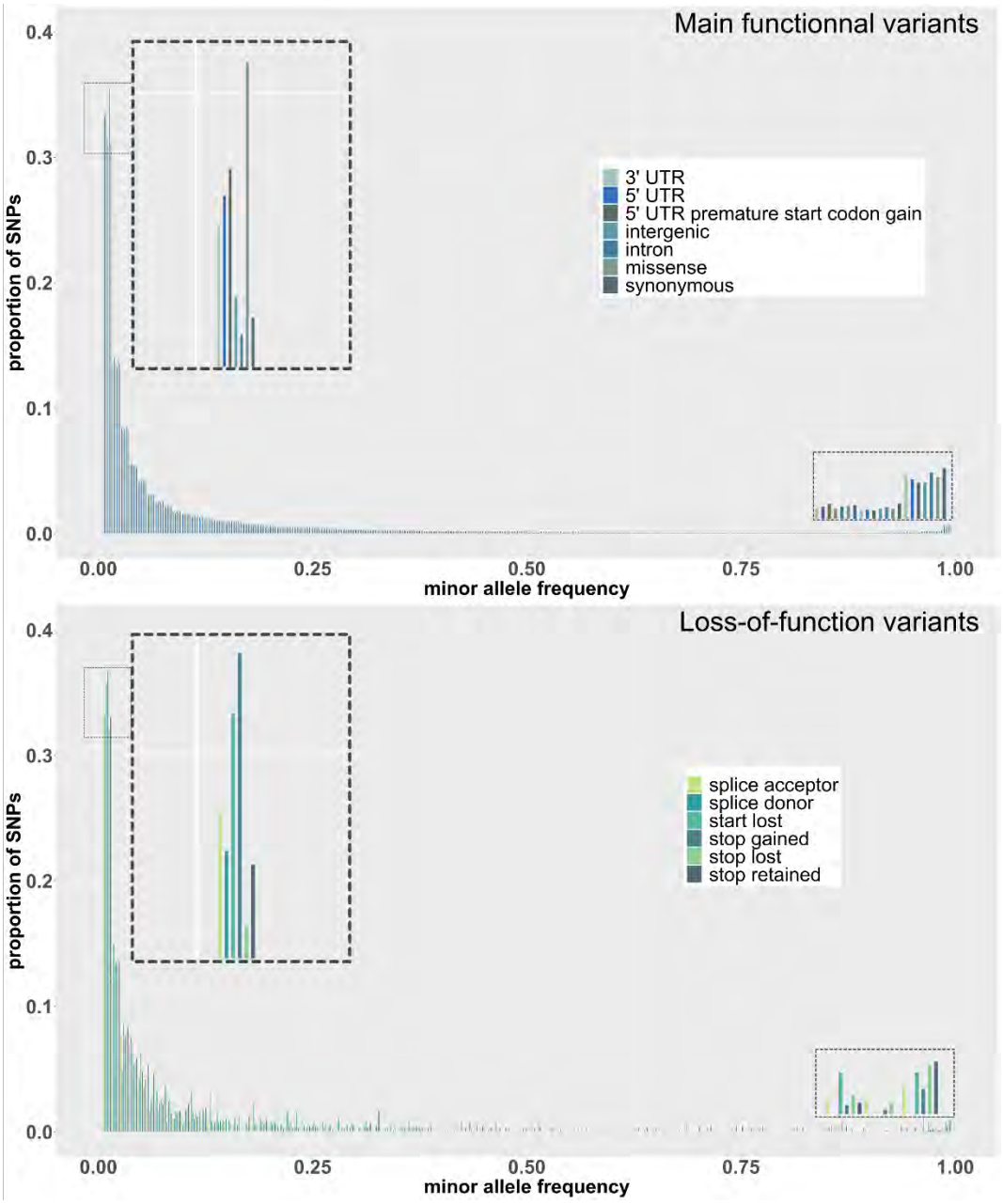


Figure 23: Allele frequency spectrum for different categories of SNP, based on derived allele frequency. Taking into account the derived allele frequency instead of the minor allele frequency allows to observe the derived alleles that appeared in the ruderalis subspecies and reached fixation in this subspecies (potentially because of a strong positive selection). On upper panel, 1 367 612 main functional variants (696 823 intergenic, 328 780 in introns, 73 703 missense, 62 875 synonymous, 102 885 3'_UTR, 88 321 5'_UTR and 14 225 5'_UTR_premature_start_codon_gain), were used to construct a site frequency spectrum. On the lower panel, 4 783 loss of function variant (2 418 stop gained, 1 280 stop lost, 405 stop retained, 368 splice acceptor, 186 splice donor and 126 start lost) were used to construct another site frequency spectrum on rare functional variants.

b) Genes under selective pressures in *M. polymorpha*

Once the *D*, *H* and *E* statistics were calculated for all the *M. polymorpha* ssp. *ruderalis* genes, we looked at the genes with the most extreme selection signature. To do so, we considered a dataset of 18 140 genes that contained at least 4 SNPs from the folded dataset on which the Tajima's *D* statistics was calculated. The top 10% of genes (i.e., 1 814) with the highest values of Tajima's *D* were selected as a gene set with the most pronounced signature of balancing selection. Among the remaining 16 326 genes, we kept 11 932 genes for which the ancestral/derived alleles identification could be inferred for at least 50% of their SNPs. This gene set allowed to select 777 genes with marked signature of background selection based on a Zeng's *E* values < 10% genome-wide (i.e. on the 18 140 genes) quantile value. Finally, the same gene set allowed to select 1 374 genes with marked signatures of soft or hard selective sweep based on Fay and Wu's *H* values < 10% or *E* values > 90% genome-wide (i.e. on 18 140 genes) quantile value. Hence, 3 965 genes were considered has having pronounced selection signatures based on the *D*, *H* and *E* neutrality test statistics.

These lists of genes were then used to perform functional enrichment test, by considering the occurrence of Interpro (IPR) domains in the three gene categories (balancing, background and sweep selection) and comparing it to IPR domain occurrence in the whole reference genome, with a hypergeometric test. The IPR domains with a q-value (FDR multiple test correction) inferior to 0.05 were considered as significantly enriched in the genes under a given selective force. The functions enriched in genes under background selection were mostly linked to basic cell functions dealing with DNA or proteins (Figure 24, see SupData1.4 for full table). For instance, there was genes linked to cell division, like the meiotic recombination protein Dmc1 (IPR011940) that is essential for recombination and double strand repair during meiosis **(H. Chen et al., 2021)**, or the tubulins (IPR023123) that allow to segregate the genetic material during mitosis. Many gene-regulating proteins are also conserved like C2H2 zing fingers and B3 transcription factors, as well as F-box domain proteins among which MpFBW2 involved in RNA silencing **(Hacquard et al., 2022)**. Two glycosyl hydrolase family 16 proteins, that are involved in cell wall modulation are also conserved.

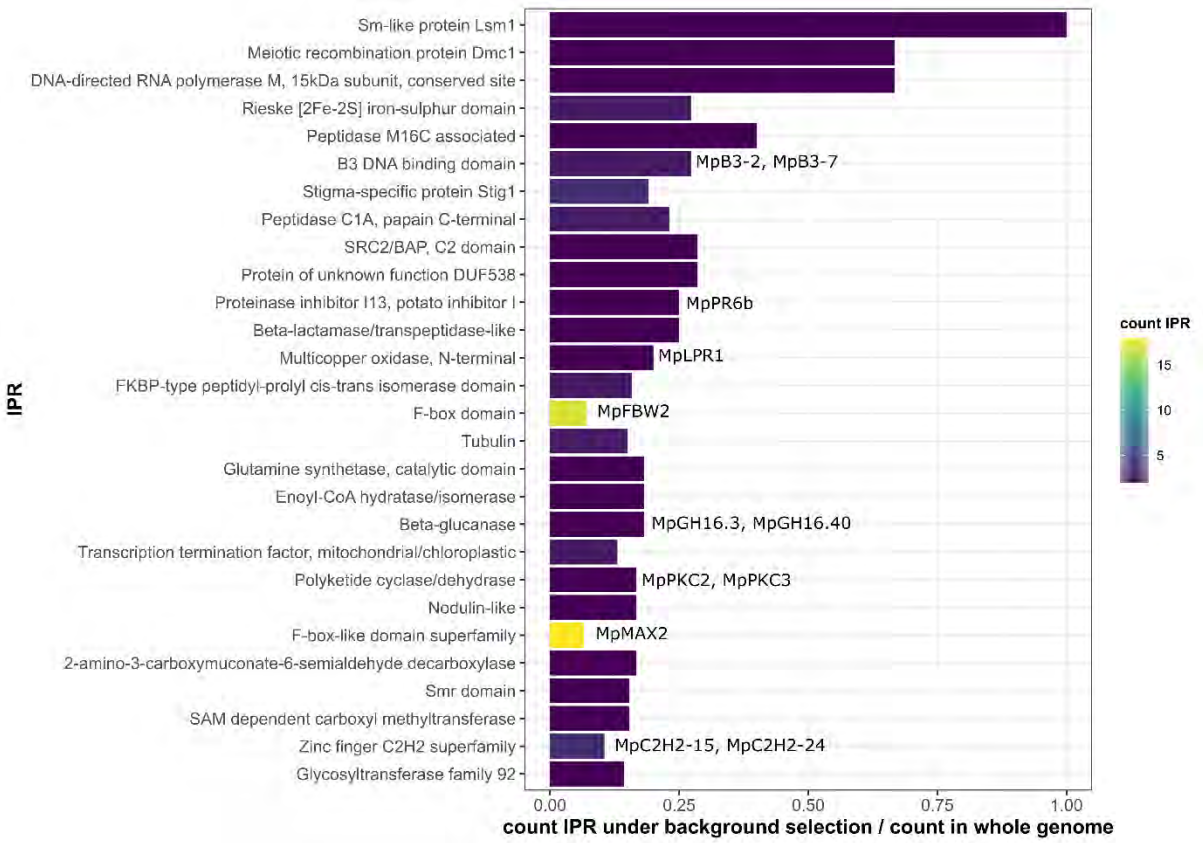


Figure 24: Barplot representing the IPR domains enriched in M. polymorpha's genes under background selection. The IPR list was refined manually to suppress domains names referring to the same genes and was sorted in ascending order of p-value. When available, genes symbols are specified.

Among the genes undergoing selective sweep there are nine glycosyl hydrolase 16, two transcription factor of the MADS-box family, two chaperones (HSP90), and two pathogenesis related proteins (PR6 family) (SupData1.5). Families under selective sweep seem to have more adaptive roles than families under the other type of directional selection (background selection).

Many of the genes under balancing selection are large families of oxidation-reduction enzymes (Figure 25, see SupData1.6 for full table) involved in various biochemical pathways that have roles in plant defense: the peroxidases (31 genes under balancing selection), the catalases (2 genes under balancing selection), the lipoxygenases (9 genes under balancing selection), the polyphenol oxydases (14 genes under balancing selection), and the cupredoxins (5 genes under balancing selection). Other families under balancing selection also seem involved in plant defense, like the concavalin-A lectin family, that contains GH16 and RLK-Pelle proteins (19 genes under balancing selection), the Bulb type lectin (with some GH17 genes and a total of 8 genes under balancing selection) or the Ricin B lectins (3 genes under balancing selection).

These carbohydrate binding proteins are involved in a variety of biological processes in the plant, among which defense against pathogens and predators, nitrogen storage, regulation of plant growth and development **(De Coninck & Van Damme, 2021)**. There are also 8 PR-1 (pathogenesis related 1) proteins that are involved in response to pathogen with their antimicrobial activity and defense signal amplification **(Breen et al., 2017)**, or 3 decaprenyl diphosphate synthase (MpCPT2, 5 and 7) that play a role in the synthesis of polyprenyl pyrophosphates, the precursors of the terpenoid pathway.

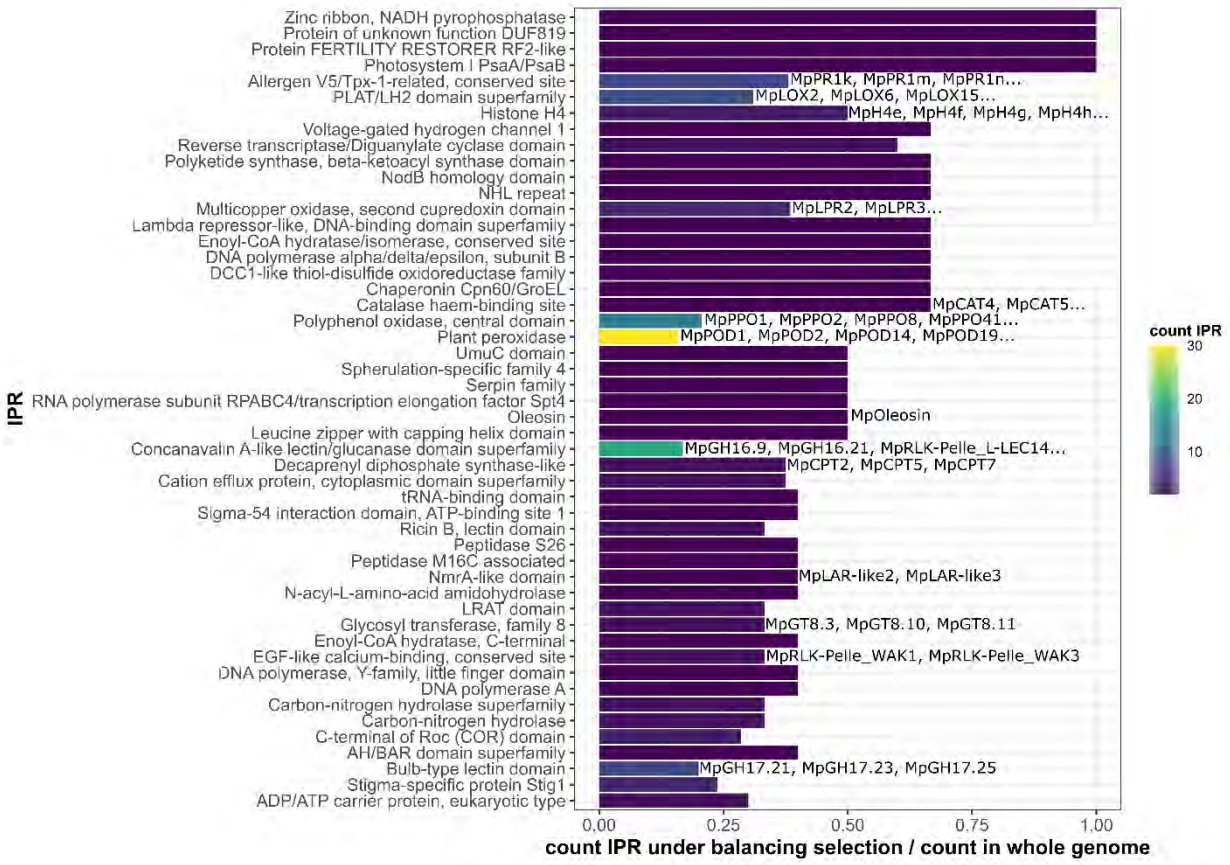


Figure 25: Barplot representing the IPR domains enriched in M. polymorpha's genes under balancing selection. The IPR list was refined manually to suppress domains names referring to the same genes and was sorted in ascending order of p-value. When available, genes symbols are specified.

5) Investigation of interspecific comparisons of intraspecific selection signatures in land plant species

We then crossed the selection signature on *M. polymorpha* genes with the results obtained in *M. truncatula* and *A. thaliana*. Since the three species have diverged a long time ago, the gene lists were determined by looking at the top 20% genes with extreme *D*, *H* and *E* value, with the exact same steps described for the top 10% genes in *M. polymorpha*. The correspondence

between genes in the three species was made based on orthogroups reconstructed using OrthoFinder v2.5.2 (Emms & Kelly, 2019). To ensure an accurate reconstruction of orthogroups, 36 species covering the main lineages of land plants were added to the above-mentioned species (SupData1.7 sheet 1 and 2). A first round of OrthoFinder was performed and a careful inspection of the estimated species phylogeny has been conducted to ensure the absence of errors compared to the phylogeny of land plants (One Thousand Plant Transcriptomes Initiative, 2019; The Angiosperm Phylogeny Group, 2016). A consistent tree has been reconstructed and the tree has been used for a second run of OrthoFinder with the option "msa", enabling to reconstruct the orthogroups using alignment and phylogenetic reconstruction methods. We then focused on hierarchical orthogroups (HOG, i.e. groups of genes containing all orthologs and in paralogs duplicated after the speciation of the last common ancestor of the species studied) showing at least one gene in each of the three species under the same selective regime (SupData1.7, sheet 4). Among the orthogroups with genes under background selection in the three species, can be cited the orthogroup containing the MpFERONIA receptor like kinase that is crucial for *Marchantia* and angiosperm development, primarily linked to cell expansion and maintenance of cellular integrity (Galindo-Trigo et al., 2016; Malivert & Hamant, 2023; Mecchia et al., 2022). This gene is present in only one copy in *M. polymorpha* genome, but has been duplicated in the angiosperms: there is 14 copies in *A. thaliana* and 22 copies in *M. truncatula*. In these two species the background selection has been relaxed on the multiple copies, with only one still being under background selection in both species. Other orthogroups with shared background selection contain transcription factors such as the WRKY transcription factors (WRK9 in *M. polymorpha*, 7 genes in *A. thaliana* among which WRK27 (AT5G52830) and WRKY65 (AT1G29280) are under background selection, and 10 genes in *M. truncatula* with only one in background selection (MtrunA17Chr8g0379461)) or the MYB transcription factors (Mp1R-MYB10 in *M. polymorpha*, 2 genes in *A. thaliana*, among which ADA2a (AT3G07740) is under balancing selection and 8 genes in *M. truncatula* with only 1 under background selection). Orthogroups with genes sharing a common selective sweep signature count aluminium activated malate transporters that promote aluminium tolerance (genes under swep: 1/2 genes in *M. polymorpha*, 2/5 genes in *A. thaliana* and 1/6 genes in *M. truncatula*) and auxin response factors (ARF) that regulate the expression of genes in response to auxin signalling (genes under sweep: 1/1 gene *M. polymorpha*, 2/14 genes in *A. thaliana* and 2/7 genes in *M. truncatula*). Finally, among the genes

under similar balancing selection in the three species, we can find peroxidases, that are one of the rare *M. polymorpha* highly duplicated gene family (genes under balancing selection: 11/41 genes *M. polymorpha*, 1/8 genes in *A. thaliana* and 1/4 genes in *M. truncatula*), dirigent proteins that are involved in lignin biosynthesis **(Y.-Q. Gao et al., 2023)** (1/4 genes *M. polymorpha*, 3/14 genes in *A. thaliana* and 7/25 genes in *M. truncatula* under balancing selection), pathogenesis related 5 proteins (1/1 *M. polymorpha*, 4/21 *A. thaliana*, 10/40 *M. truncatula*) or terpenoid synthases (1/2 *M. polymorpha*, 2/2 *A. thaliana*, 1/7 *M. truncatula*). A lot of genes under shared balancing selection between the three species are therefore genes involved in plant stress response. This is not surprising since this tendency has already been observed in plants **(Karasov et al., 2014)** and in animal species **(Croze et al., 2016)**, with an emphasis on the polymorphism of immunity genes. This conservation of balancing selection in immune genes is probably due to the co-evolution between the host and the parasite, in which both parties try to escape the invading or defending mechanisms of the other. Balancing selection on immunity genes can be due to the "trench warfare" **(Stahl et al., 1999)** mechanism of coevolution that results from fluctuation of the resistance allele frequency along time and space, because of a cost-benefit balance (the costs being, for instance, a deficit in development coming from an antagonistic crosstalk of developmental and defensive hormonal pathways **(Karasov et al., 2017)**). This is opposed to the notion of co-evolutionary "arms race" that leads to the increase of the disease resistance allele frequency in the population, until it is by-passed by the pathogen.

III) Concluding remarks on chapter 1

This chapter has presented the bases of the genomic resources that have been developed for the model liverwort *Marchantia polymorpha*. Our collection of 135 accessions was sequenced and the polymorphic positions compared to the reference genome were called, leading to a high-density SNP dataset. Since SNPs are currently one of the bases of many genomic analysis, like population genomics or quantitative genetics, this dataset enables us to explore the genetic diversity in *Marchantia polymorpha*. Even though this dataset was not designed to answer to phylogeographic questions (the sampling does not cover homogeneously a delimited geographic range), it still enabled us to detect the weak correlation between the genetic structuration of our population and its geographic distribution. This suggests frequent outcrossing and long-distance gene flow in *Marchantia polymorpha*. The exact mechanisms of

these frequent and long-distance genetic exchanges are still unknown but could be linked to the transport of Marchantia via horticultural trade (Marble et al., 2017), and to an ability of the bryophyte to disperse its gametes on long distances (Sandler et al., 2023).

The site frequency spectrum of the SNP resource was examined, indicating general tendencies for selective pressures associated with variant types, that do not seem to deviate from patterns found in angiosperms, and many other organisms. For instance, strong purifying selection on missense variant seem to be a general rule. On the other hand, it also seems that loss of function variants can also be associated with adaptive processes through balancing selection or selective sweep mechanisms (Monroe et al., 2021). Summary statistics (Tajima's *D*, Fay and Wu's *H* and Zeng's *E*) pinpointing the type of selection acting on genomic regions were also calculated based on the SFS. The calculation steps were optimised in order to correct for the presence of missing data at polymorphic sites, and to obtain a more flexible method of calculation. Here, only genes have been considered for the calculation of these estimators, but this method of calculation could also be used on functional domains in genes, or on regulatory regions of genes. This would allow to go further in the understanding of the selective forces at work in specific genes of interest, or to explore how the regulatory regions are impacted by selection at a whole genome scale.

The exploration conducted here on *M. polymorpha's* genes revealed the nature of its conserved genes under purifying selection (genes linked to basic cell functions) and of its gene under other type of selection. The genes under selective sweep (selection of a mutated haplotype that likely appeared in *M. polymorpha* ssp. *ruderalis*) had more adaptive roles, like the two pathogenesis related proteins, or the multiple glycosyl hydralase family 16. Quite logically, these types of proteins were also found in the list of genes harbouring a derived missense variant almost fixed in the population. The genes with a balancing selection signature were also large gene families potentially playing a part in the plant's adaptation. For instance, there were an important number of class III peroxidases, that take part in a broad range of stress responses and developmental processes (Kidwai et al., 2020), as well as polyphenol oxidases that are involved in the biosynthesis of defensive metabolites (J. Zhang & Sun, 2021).

These results were then compared with the genes under selection in two other plant species: *A. thaliana* and *M. truncatula*. This led to some candidate families with shared selection signature in the three species, like the receptor-like kinase family in which the FERONIA gene

is, or the class III peroxidase family. This comparative dataset was sometimes difficult to navigate because of the duplications that asymmetrically occurred in the three species lineages, leading to a relaxation of selective constraints on most of the duplicated copies. It is nevertheless an interesting resource to compare genes of interest in different species, and to see if they share the same selection signatures or not.

Chapter II

Genome-wide association studies to understand *Marchantia polymorpha's* adaptation to biotic and abiotic stresses

As it was mentioned in the introductive chapter, a way of exploring the adaptative strategies implemented in plant genomes is to link genomic variation with the environmental conditions or the plant's phenotype. We therefore performed genome wide association studies on our dataset of *M. polymorpha* ssp. *ruderalis* to find genetic elements of its adaptation to abiotic and biotic stresses. The first approach of genetic association explored with our dataset of *M. polymorpha* was linked to climatic conditions.

I) Genome environment association study in *M. polymorpha*

1) Preparation of the data and GEA implementation

a) Retrieval of the climatic data for *M. polymorpha* ssp. *ruderalis* accessions

Most of the accessions of our collection have GPS coordinates for their sampling site (127 accessions out of the 135). These coordinates are either the precise information recorded during the sampling, or, for accessions from eastern Europe and Scandinavia, an inference I made based on information about the villages and countries where the plants where sampled. These two categories are specified on the information table (SupData1.1). With this GPS coordinates, information about the sampling site could be retrieved. The altitude of the plants was inferred with the online tool (https://www.gpsvisualizer.com/elevation), and climatic information is available on the second version of the Worldclim database, that aggregates climate data from 1970 to 2000 **(Fick & Hijmans, 2017)**. Information from the worldclim tif files "wc2.1_30s_bio/prec/srad/vapr" were extracted for each spatial point in R with the sp package. The extraction was performed on the 127 *M. polymorpha* accession for which GPS coordinates were available, but failed for two of them. The variables retrieved were of two sorts: 19 bioclimatic variables derived from the temperature and precipitation record on 12 months (BIO1 to BIO19), and monthly climate data for the water vapor pressure (vapr kPa), the solar radiation (srad KJ m^{-2} day^{-1}) and the precipitations (prec mm).

For the latter, principal component analysis (PCA) were performed for each of the climatic variable on the 12 months' data, and only the two first principal components (PC) of each variable were analysed. These two components allowed to describe 87,85% of variance for the precipitation, 98.44% of variance for the solar radiation and 96.45% of variance for the water vapour pressure.

Since most of the climatic variables are correlated, PCA was also performed on the 19 bioclimatic variables, separating the temperature linked variables (BIO1 to BIO11) and the precipitation linked variables (BIO12 to BIO19). The first 3 principal components of each analysis were retained, accounting for 98.37% of the variance of the precipitation linked variables, and 95.02% of the temperature linked variables. The PCA analysis on these variables allow to visualise heterogeneity in the climatic environment of our *Marchantia* collection. For the two categories of variables (temperature and precipitation linked), the PC1 is clearly constructed based on the annual values of the climate, while the PC2 is based on the intensity of climate variation along the year. On the temperature graph, the Scandinavian accessions (in yellow) are subjected to cold temperatures and strong temperature seasonality, the British ones to medium temperature and very small variation along the year and the Pyreneans to higher temperature and variation than the British ones. Accessions from the US and Japan show extremely different climatic conditions compare to the Europeans (Figures 26 & 27). According to the PCA on precipitation variables, the separation between the geographic groups is less clear, with all accessions having similar precipitation variation along the year (except from the ones from the West Coast of the US). As someone from the south of France I was surprized to realize that England is not always the place where it rains the most.

The geographical groups (Pyrenean accessions, British accessions) clearly cluster on the PCA based on the environmental variables, which is logic, but may be intensified by the fact that the sampling is not homogenous Some accessions from these groups have different climatic conditions from the rest of their clusters (Lakedist for the precipitations, for instance), and our analysis showed that the geographical distribution of the population did not exhibit substantial correlation with its genetic structure. Therefore, this genome environment association study should be able to identify candidates associated with the differential environmental conditions.

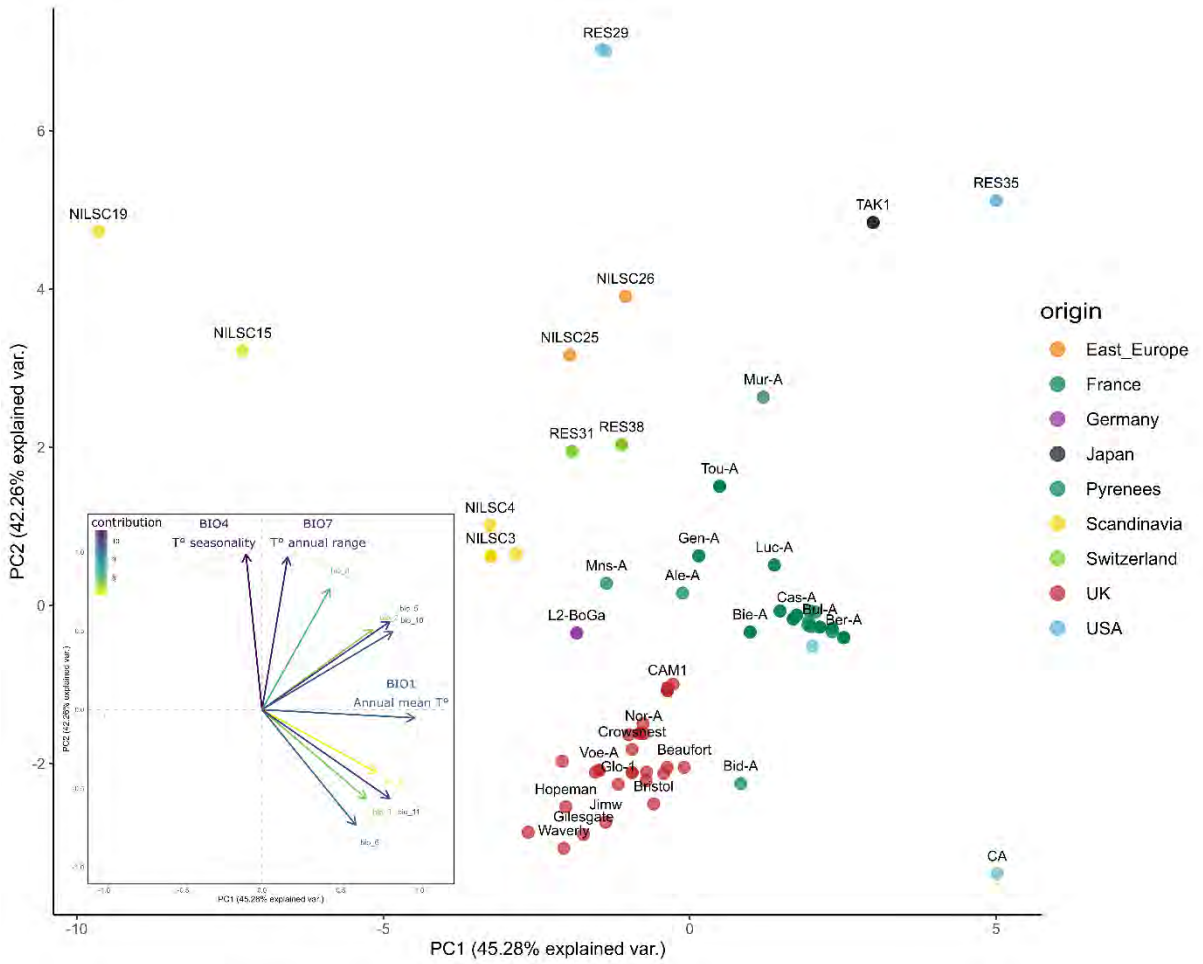


Figure 26: Distribution of the accessions along the PC1 and PC2 of temperature linked bioclimatic variables, and contribution of the variables to the construction of the PC.

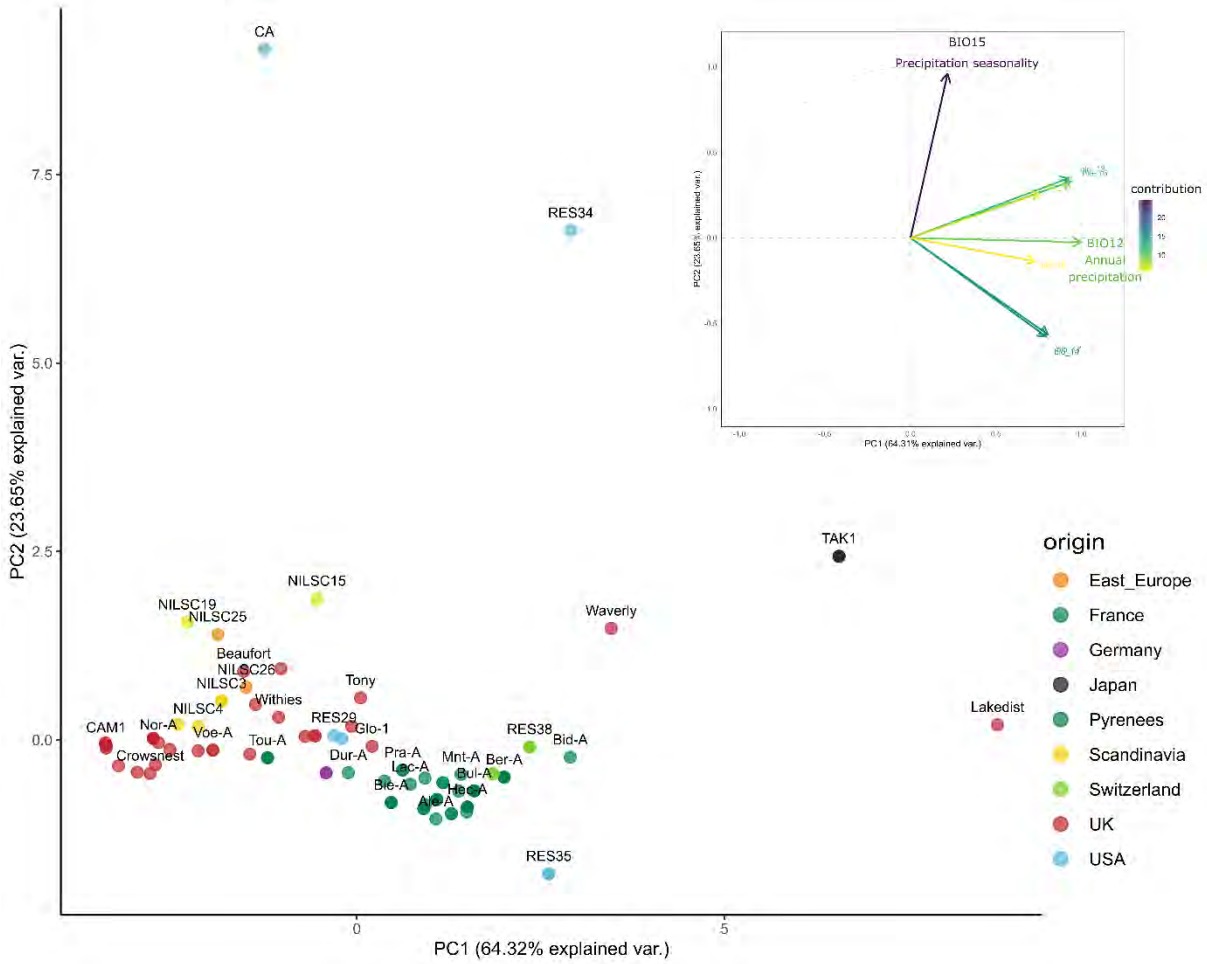


Figure 27: Distribution of the accessions along the PC1 and PC2 of the precipitation linked bioclimatic variables, and contribution of the variables to the construction of the PC.

A trial of dimension reduction was also made on all bioclimatic variables and all monthly information of the vapr, srad and prec climatic data. The 4 first PCs were retained for GWAS analysis, representing 89.63% of the variables' variance.

Since the population structure is strong between the subspecies and that the sampling sizes in the three subspecies is not homogenous, we carried out the analysis on the subset of 96 *ruderalis* accessions.

b) GEA analysis

The SNP data for the 96 accessions was extracted from the unfiltered VCF file, to perform the independent association analysis on all sites and then allow to test differential filters of minor allele frequency MAF and missing data. The GWAS analysis was performed using the linear mixed model implemented in the GEMMA software **(X. Zhou & Stephens, 2012)**. Population structure was accounted for with a centred relatedness matrix computed with the software **(-**

`gk 1 option`). Each bioclimatic variable and principal component mentioned previously was tested separately in the univariate linear mixed model (with option `-lmm 4` (performs Wald test, score test and likelihood ratio test) and `-maf 0 -miss 1` to get the association result for all variants).

The results of the GEMMA run on each SNP were then processed with the local score **(Bonhomme et al., 2019; Fariello et al., 2017)**, after filtering for a MAF of 0.05 and a maximum of 15 accessions with missing data per site. A ξ value of 2 was chosen, meaning that all SNP with a p-value below 10^{-2} contribute to the increase of the Lindley process. Chromosome-specific significance thresholds (α = 5%) were estimated using a resampling which allow to estimate the coefficients of the Gumble distribution used to model the local score under the assumption of no association **(Bonhomme et al., 2019; Fariello et al., 2017)**. The R scripts used to compute the local score and significance thresholds are available at https://forge-dga.jouy.inra.fr/projects/local-score/documents. All the significant local score peaks coordinates were extracted, and they were annotated with their overlapping (when existing), downstream and upstream genes.

2) Details on the candidate regions found

a) General description of the results

The GWAS analysis was performed independently on 36 environmental variables (Table 2), leading to the identification of 64 peaks, 36 of which are common between two or more variables. This was expected since most of these variables are strongly correlated. Some variables (like PC2 and 3 of precipitation bioclimatic variables) show higher p-values than expected (Appendix B and C).

Table 2: Detail of all the variable tested in the genome environment association study, and of the number of significant genomic regions associated

variable	description	Number of significant peaks
Precip_PC1	PC1 of PCA on BIO12 to 19	4
Precip_PC2	PC2 of PCA on BIO12 to 19	7
Precip_PC3	PC3 of PCA on BIO12 to 19	4
Temp_PC1	PC1 of PCA on BIO1 to 11	3
Temp_PC2	PC2 of PCA on BIO1 to 11	4
Temp_PC3	PC3 of PCA on BIO1 to 11	2

Altitude	Heigh above sea level (m)	4
Bio_1	Annual mean temperature	4
Bio_2	Mean diurnal range	5
Bio_3	Isothermality (Bio_2/Bio7)(x100)	3
Bio_4	Temperature seasonality (standard deviation x100)	4
Bio_5	Max temperature of warmest month	8
Bio_6	Min temperature of coldest month	2
Bio_7	Temperature annual range (Bio_5-Bio_6)	6
Bio_8	Mean temperature of wettest quarter	3
Bio_9	Mean temperature of driest quarter	2
Bio_10	Mean temperature of warmest quarter	6
Bio_11	Mean temperature of coldest quarter	4
Bio_12	Annual precipitation	5
Bio_13	Precipitation of wettest month	6
Bio_14	Precipitation of driest month	3
Bio_15	Precipitation seasonality	6
Bio_16	Precipitation of wettest quarter	4
Bio_17	Precipitation of driest quarter	5
Bio_18	Precipitation of warmest quarter	4
Bio_19	Precipitation of coldest quarter	4
Prec_dim1	PC1 of PCA on monthly precipitations	3
Prec_dim2	PC2 of PCA on monthly precipitations	3
Srad_dim1	PC1 of PCA on monthly solar radiation	8
Srad_dim2	PC2 of PCA on monthly solar radiation	5
Vapr_dim1	PC1 of PCA on monthly water vapor pressure	3
Vapr_dim2	PC2 of PCA on monthly water vapor pressure	4
All_var_dim1	PC1 of the PCA on bio_1 to 19 and all monthly variables	7
All_var_dim2	PC2 of the PCA on bio_1 to 19 and all monthly variables	4
All_var_dim3	PC3 of the PCA on bio_1 to 19 and all monthly variables	3
All_val_dim4	PC4 of the PCA on bio_1 to 19 and all monthly variables	3

The list of candidates was treated as a whole by doing a domain enrichment to see the general tendency of the regions found, and then some candidates were investigated more in depth (table with all candidates available: SupData2.1). These were chosen depending on their position relative to the peak, the climatic variables they were associated with and their function.

b) Enrichment on the full list of candidates

The list of upstream, downstream and overlapping genes of the 64 candidate regions was used to perform an IPR domain enrichment. This enrichment can provide insights on the most common functions in regions associated to climatic variations, regardless of which gene is the

real causal variant (which is often hard to determine without functional validation). The enrichment test was performed by comparing the occurrence of IPR terms in the GEA candidate gene list with their occurrence in the whole reference genome, with a hyper-geometric test (phyper function in R). IPR were considered as significant if their p-value corrected for multiple testing (Benjamini-Hochberg) was inferior to 0.05 and if they were present in at least two genes in the whole genome (SupData2.2). Among the big gene families in the whole genome, the ones enriched in the GEA results were families linked with response to pathogens (NLR with the NB-ARC and LRR domain, LURP-1 related), with detoxification (Glutathione S transferase), or with transport (ABC-2 type transporter, ammonium transporter).

c) Focus on some candidates

Some of the 64 regions found associated with the climatic conditions displayed interesting gene functions or association with a great diversity of climatic variables and were focused on to understand a bit more their role in *M. polymorpha's* adaptation to climatic conditions. Most of them are represented in the Miami plot of the principal components linked to temperature and precipitation (Figure 28).

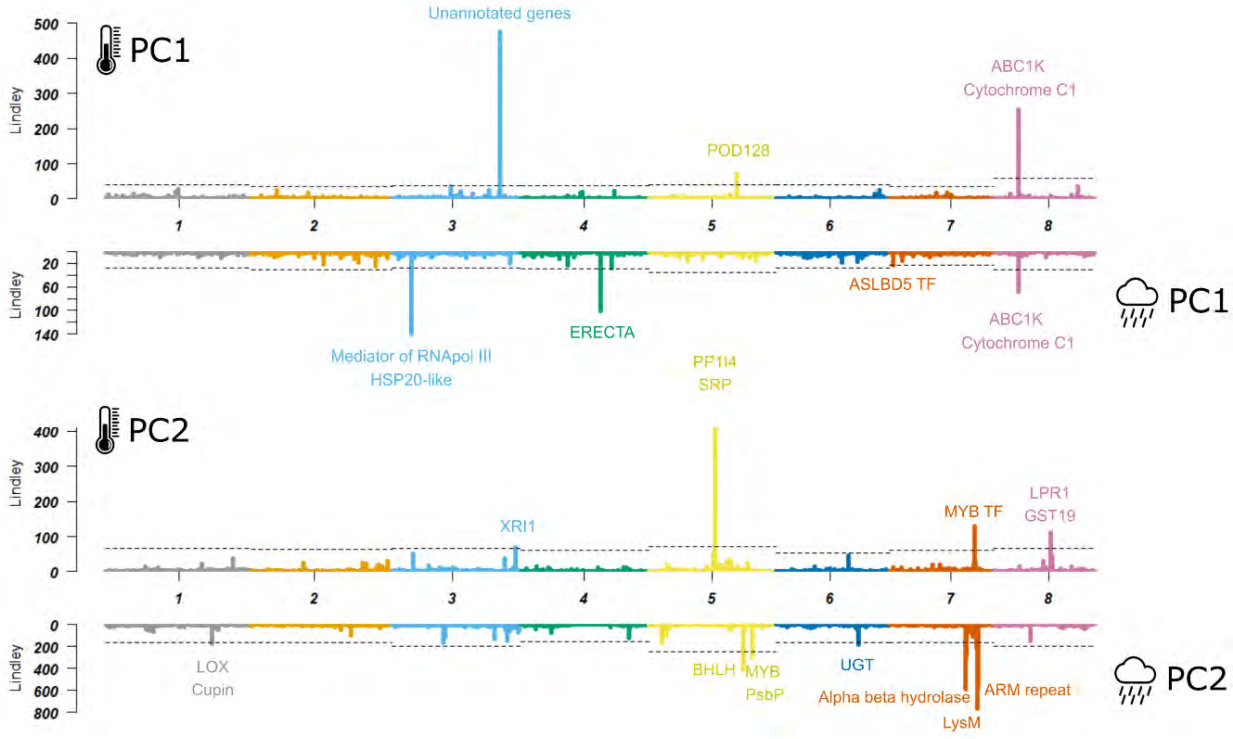


Figure 28: Genomic regions in M. polymorpha associated with the 2 first principal components of the PCA on temperature and precipitation bioclimatic variables.

The ABC1K candidate

The main QTL that seems linked to temperature and precipitations variables is the one on chromosome 8. It is significantly associated with 11 variables, some but not all of them being correlated (Temp_PC1, Precip_PC1, all_var_dim1, prec_dim1, bio_1, bio_2, bio_5, bio_10, bio_12, bio_17 and srad_dim). This peak is surrounded by an atypical protein kinase **ABC1K** (activity of bc1 complex kinase Mp8g04680) and a **cytochrome c1** protein (Mp8g04690). The significant region is upstream of both genes but is a bit closer to the ABC1K, potentially meaning that the causal variant is in its regulatory region. The prevalence of the alternative haplotype responsible for the association signal is quite important in the population (24/96 accessions), and is associated with high values of precipitation, temperature and solar radiation (Figure 29 & Figure 30).

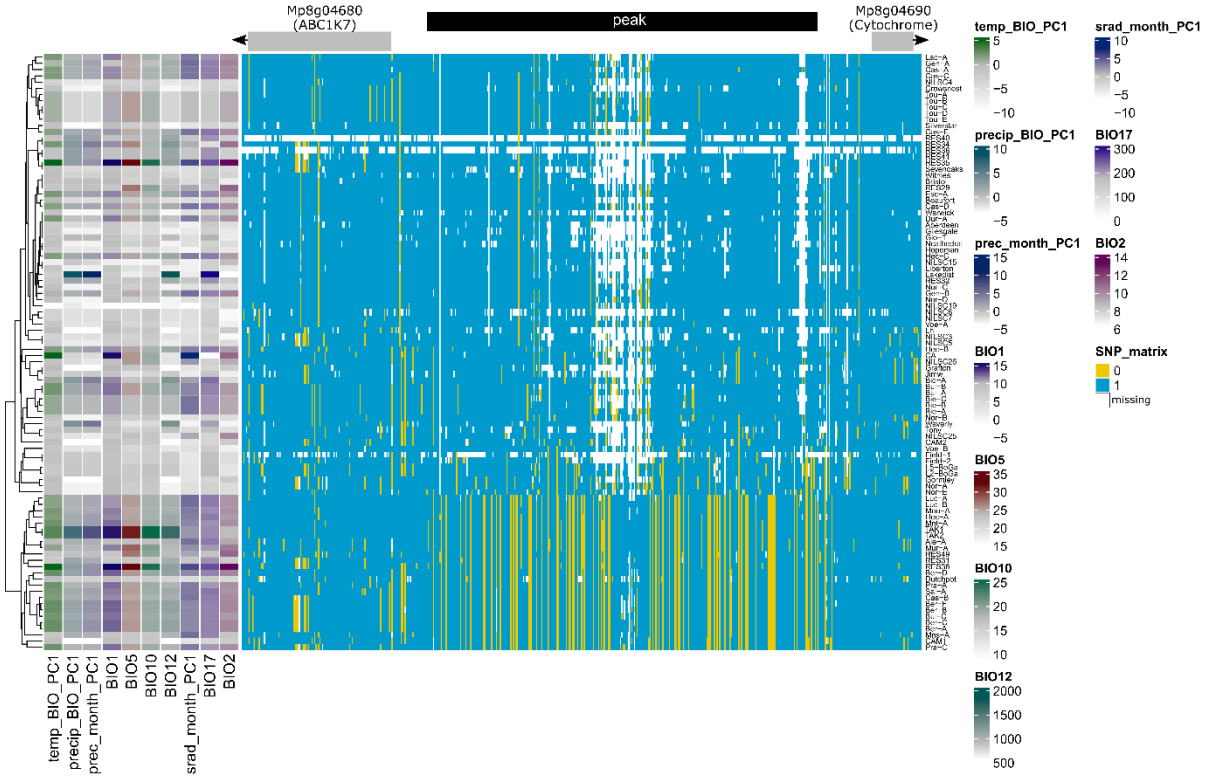


Figure 29: Heatmap representing the SNP matrix in the region of the GEA peak surrounded by the ABC1K and the cytochrome c1. The accessions are clustered according to SNP allele similarity, and annotated with the climatic conditions with which the peak is associated.

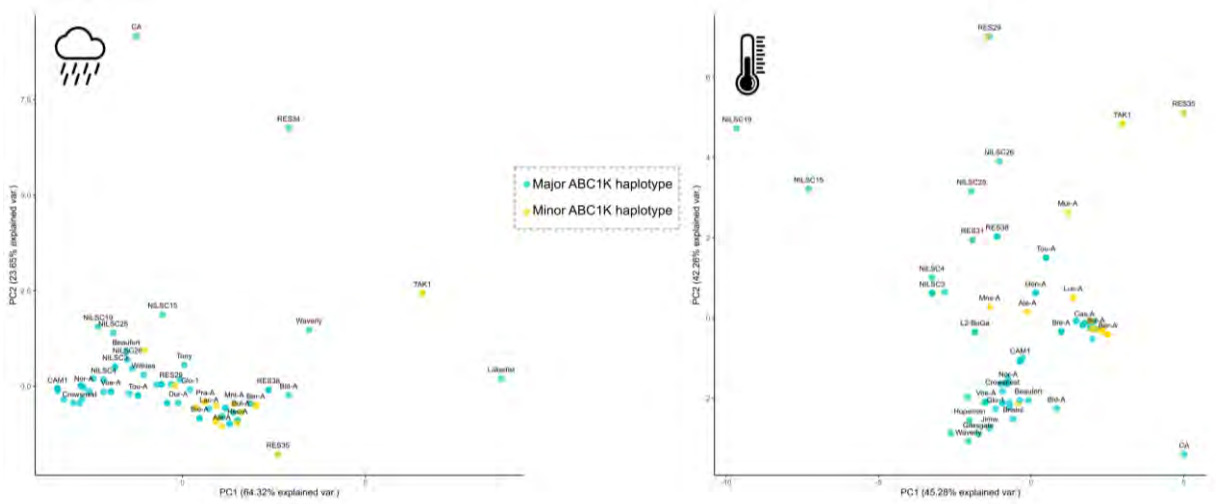


Figure 30: Distribution of the accessions with the minor (yellow) and major (blue) haplotype in the ABC1K genomic region. The distribution is represented on the space formed by PC1 and PC2 of the PCA on precipitation linked variables (left) and PC1 and PC2 of the PCA on temperature-linked variables (right). Most of the accessions with the minor phenotype come from sampling sites with high annual mean temperature (PC1 temperature) and high annual mean precipitations (PC1 precipitations).

ABC1K is a family of protein kinases conserved in a variety of species of archaea bacteria and eukaryotes, but that specifically expanded in plants, forming a clade that is specific of photosynthetic organisms. In flowering plants, these kinases have been shown to be located in plastids or mitochondria **(Lundquist et al., 2012)**. In mitochondria they are involved in respiratory pathways via coenzyme Q synthesis, whereas in chloroplasts, they participate in prenylquinone synthesis and stress responses **(Manara et al., 2015)**.

To explore more specifically the potential role of this protein, a phylogeny of its orthologs in plants was made, based on a curated database containing 37 representative species (Appendix D). The orthologs were searched for with a blastp with a e-value threshold of 10^{-30} and a maximum of 4000 target sequences possible, leading to 450 hits, that were aligned with muscle5 and trimmed with trimAl to discard positions with more than 60% of gaps. A phylogenetic tree was then computed based on this alignement, using IQtree **(`-alrt 10000 -B 10000 -T AUTO`**, the Q.pfam+F+R10 model was chosen) **(Minh et al., 2020)** and visualised on ITOL **(Letunic & Bork, 2021)** (Figure 31).

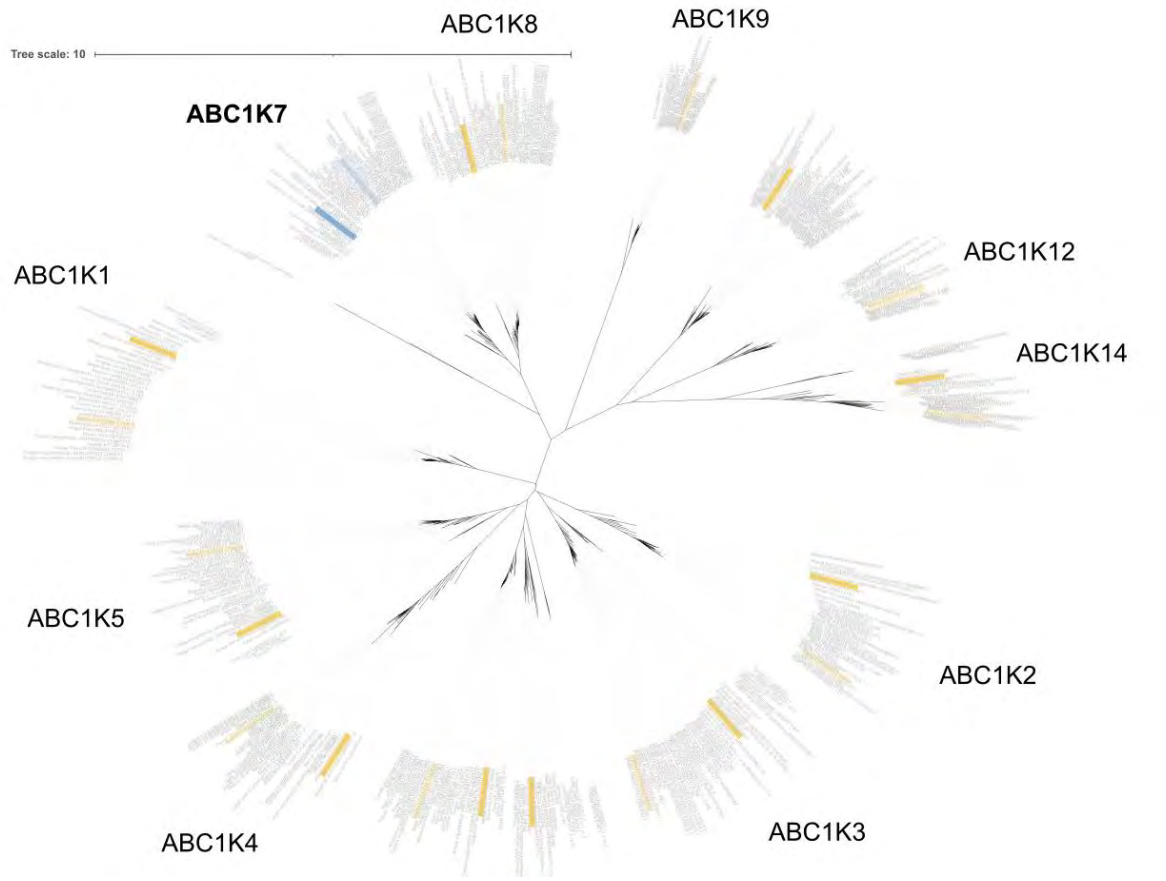


Figure 31: Orthologs of the ABC1K. The GEA candidate is highlighted in blue, and the other ABC1K are in yellow. M. polymorpha's genes are highlighted in darker shades, and A. thaliana's in lighter shades of blue (ABC1K7 family) and yellow (other ABC1K families). The ABC1K families 1 to 8 are the ones specific to the photosynthetic clade, ABC1K9 belong to the ancestral clade, and ABC1K12 and 14 to the mitochondrial clade ***(Lundquist et al., 2012)****. There are some clade specific variations of ABC1K orthologs, but the ABC1K types are overall well conserved in plants.*

It appears that our GWAS candidate is a ABC1K7, a type of ABC1K that is conserved in all plants, including algae (Appendix E). In Arabidopsis thaliana this gene is localised in the plastids and more specifically in their lipid containing organelles: the plastoglobules, and acts in concert with its close relative ABC1K8 (that is located in the inner plastid envelope) to moderate various processes. The oxidative stress response is one of them, since the plants mutated for these genes display a basal amount of oxidative stress even under normal conditions, and have a lower tolerance to ROS. They are also involved in iron homeostasis in the chloroplast, allowing the corrected distribution of iron containing proteins on the photosynthetic complexes **(Manara et al., 2014)**, as well as in chloroplast lipid synthesis, leading to modulation of the chloroplast membrane composition in response to stresses **(Manara et al., 2015)**. Finally, they are induced by exposition to ABA and mutant lines display deregulated ABA response, showing their involvement in ABA mediated stress response **(Manara et al., 2016)**, which makes sense

since ABA and ROS stress responses are intertwined in their signalling and the responses they trigger (S. Li et al., 2022). ABA is known in flowering plants to regulate environmental stress responses like stomatal closure under water loss, seed desiccation tolerance, and salt, desiccation and freezing tolerance. In bryophytes, it has been shown to be linked with desiccation tolerance (Khatun et al., 2023; Takezawa et al., 2011).

Taken together, this shows that ABC1K7 has a role in stress response in flowering plants, through its induction by abscisic acid, modulation of chloroplast lipidic compositions and oxidative stress mediation. This function is probably conserved in *M. polymorpha* and explains why this gene appeared associated with diverse climatic conditions.

The isoprenyl diphosphate synthase candidate

The gene MpIDS1 (Mp3g22530) is overlapped by a significant peak on chromosome 3, associated with variables linked to temperature and humidity (all_bio_PC1, prec_month_PC2, vapr_month_PC2, bio5, bio8, bio10). Isoprenyl disphosphate synthases are ubiquitous enzymes that catalyse condensation of isopentenyl diphosphate (IPP) creating terpene precursors, like the farnesyl diphosphate synthases (FDS) that synthesises the precursor of sesquiterpenes, or the geranylgeranyl diphosphate synthases (GGDPS) that creates the precursors of diterpenes (B. Singh & Sharma, 2015) (Figure 32). To clarify the role of MpIDS1, phylogeny of its orthologs was conducted in the same fashion as for ABC1K7. This revealed that our candidate is orthologous of the two farnesyl diphosphate synthase genes of *A. thaliana:* FPS1 and FPS2, that were shown to have a regulatory impact on JA pathway and abiotic stress response (Manzano et al., 2016). This gene is therefore involved in the biosynthetic pathway of sesquiterpenes, the most diverse group of terpenoids occurring in liverworts (F. Chen et al., 2018). Interestingly, specificity of bryophytes is that their sesquiterpenes are synthesised with microbial like terpenes synthases (MTPSL), which are enzymes inherited from fungi or bacteria, and not with plant terpenes synthases present in seed plants (Jia et al., 2016). In flowering plants, terpenes have been linked with defence against biotic stresses (toxic or repellent components), defence against abiotic stresses, and signalling, notably to attract insects (B. Singh & Sharma, 2015). And in bryophytes they have a role in biotic stress resistance, but also in desiccation resistance or UV tolerance (F. Chen et al., 2018).

In single cell RNAseq data (Wang et al., 2023), the MpIDS1 was shown to be highly expressed in the oil body cells (idioblastic cells). Oil bodies are specialized organelles of the bryophytes,

where the plant synthesizes and accumulates mixtures of secondary metabolites, notably terpenes but also sterols, polyketides, phenolic compounds **(Asakawa & Ludwiczuk, 2018)**. These organelles allow *Marchantia polymorpha* to deter herbivores **(Romani et al., 2020)**, and could also respond to pathogens since they contain antimicrobial compounds **(Romani et al., 2022)**. This link found between the MpIDS1 and climatic variable suggests that oil bodies could also be involved in abiotic stress responses, but this potential defensive role of oil bodies against abiotic stressors has been studied and contested for water deprivation and osmotic and salt stress **(Romani et al., 2020)**. One hypothesis could be that MpIDS1 has an influence on biotic stress tolerance linked with climatic conditions, that MpIDS1 acts on abiotic stresses via terpene synthesis outside of the oil bodies, or that the link between the oil bodies and abiotic stresses exist but has not been proven yet. In any case this association suggests the importance of terpenes in the adaptation of *M. polymorpha* to various climatic conditions.

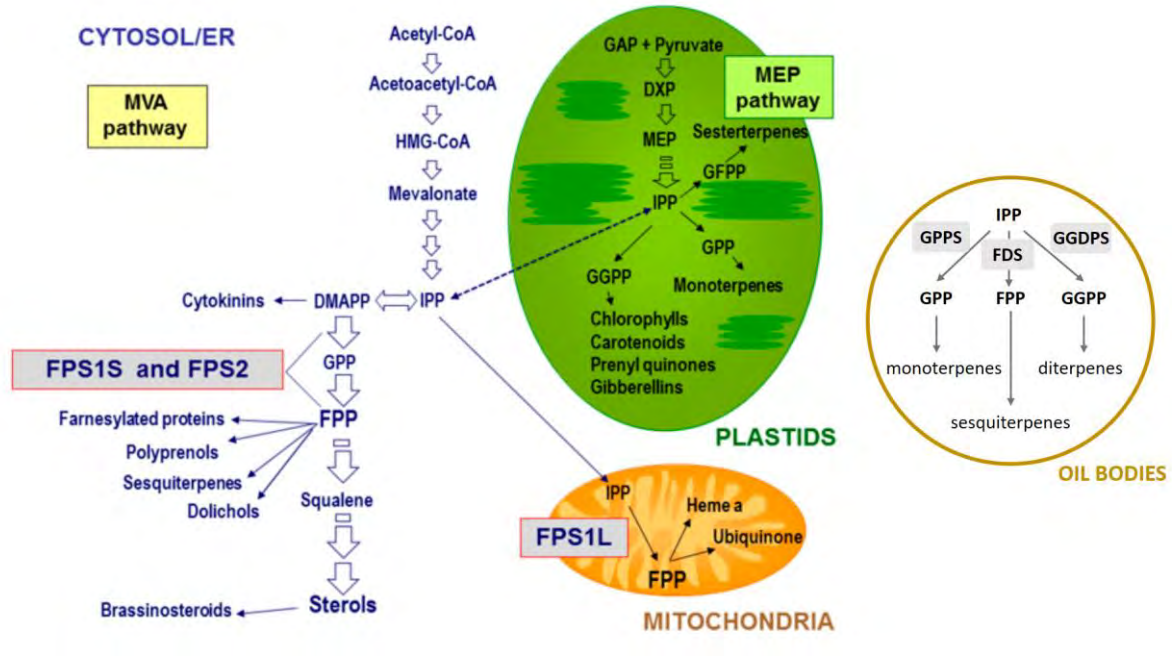


Figure 32: Overview of the isoprenoid biosynthetic pathway in plants with a focus on A. thaliana's genes (FPS1 and FPS2), adapted from ***(Manzano et al., 2016)****. The oil bodies organelles, only present in bryophytes, were added since they are an important compartment of terpene synthesis in Marchantia. FPS stands for farnesyl diphosphate synthase, that is the MpIDS1 gene in Marchantia.*

The NLR candidate

A significant region on chromosome 4, associated with the PC2 of srad and the precipitations of the wettest month, overlaps with a cluster of 2 genes related to immunity: the **NBS-LRR11** (Mp4g08790) and the **LURP-1 related** protein (Mp4g08800), known to be involved in the basal

defence of Arabidopsis thaliana against pathogens **(Baig, 2018)**. The visualisation of SNPs in the region shows that the stretch of polymorphisms causing the peak is at the tail of the NLR (potentially in the LRR domain) and between both genes (Figure 33). The alternative alleles at these polymorphisms are present in a small part of the population (the alternative allele is present in 6 accessions, except from the beginning of the peak that is associated with a minor allele present in 20 accession) and seems associated with low values of solar radiation and high values of precipitations during the wettest month.

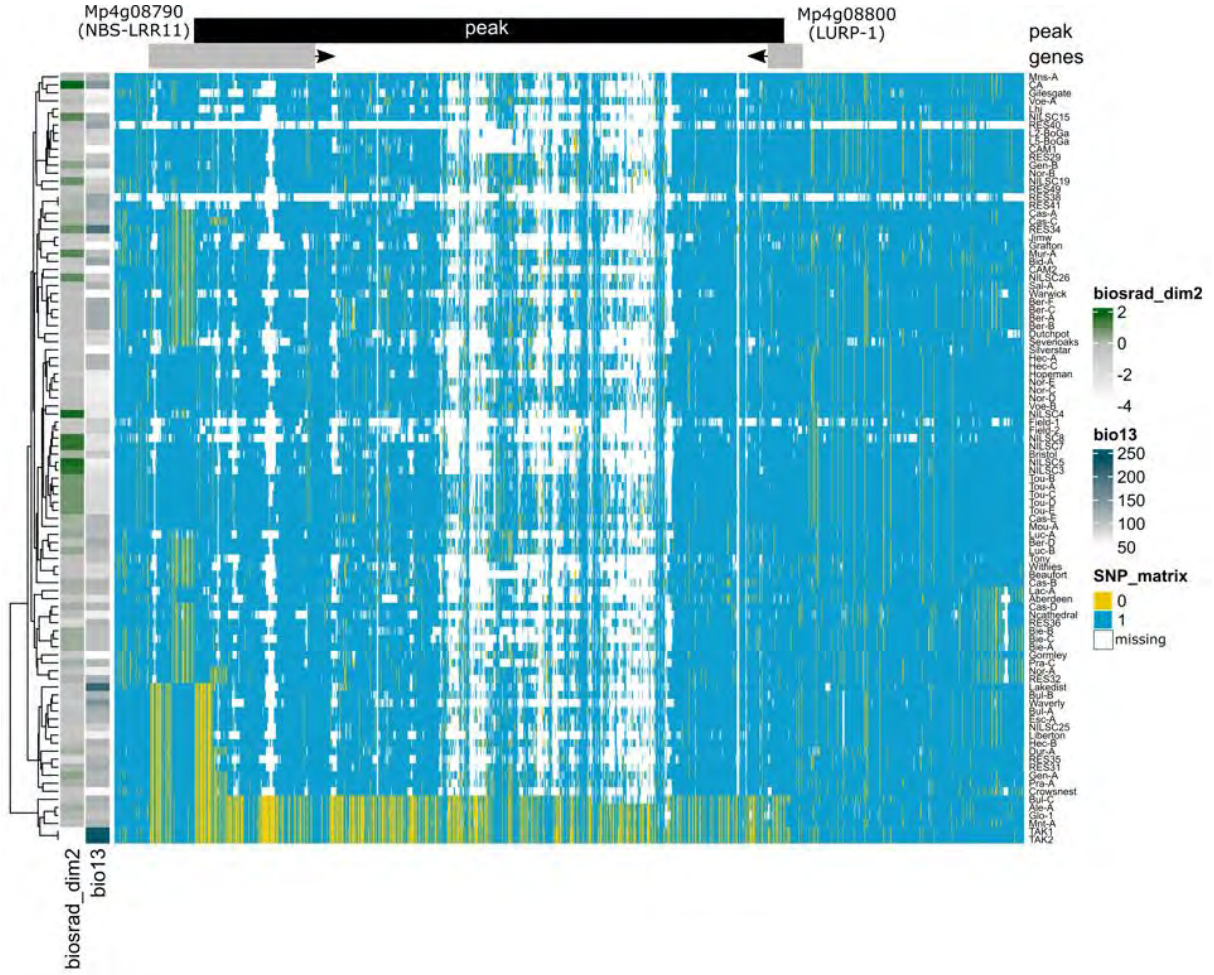


Figure 33: : Heatmap representing the SNP matrix in the region of the GEA peak surrounded by the NLR and the LURP-1 related protein. The accessions are clustered according to SNP allele similarity, and annotated with the climatic conditions with which the peak is associated.

After annotation of the reference genome with RGAugury **(P. Li et al., 2016)**, it appears that the NBS-LRR11 gene is one of the few CNL that exist in *M. polymorpha* (5 CNL annotated on a total of 39 potential NLR genes). It is one of the *M. polymorpha's* gene under balancing selection in our sample of accession (top 1.5% of the higher values of Tajima's *D*). CNL are a type of NLR that exists in all land plants and react to effector perception by forming a pore on the

membrane of the cell. Since NLR are highly diverging between plants, it is hard to guess anything more about its role without studying it in *M. polymorpha*.

The link between these genes usually involved in immunity with climatic conditions could be an indirect one (high humidity is known to promote disease outbreak **(B. K. Singh et al., 2023)**), or a direct one, meaning that biotic and abiotic stress response pathway are intertwined.

The heat shock protein (HSP) candidate

Precipitation variables (all_bio_PC1, precip_PC1, bio12, bio13, bio16) have been found associated with a region of chromosome 3, bordered upstream by a subunit of a mediator of RNA pol II (Mp3g04170) and downstream **HSP20** like chaperone (Mp3g04180). The peak is in the regulatory region of both genes, a bit closer to the mediator, that regulates the transcription by RNA polymerase II. But looking at the SNP matrix (Figure 34), it appears that the haplotype responsible for the signal continues close to the HSP but is not detected by the association test because of missing data, that may be due to the fact that TAK1 bears the alternative allele. The HSP could therefore be the gene causing the association signal.

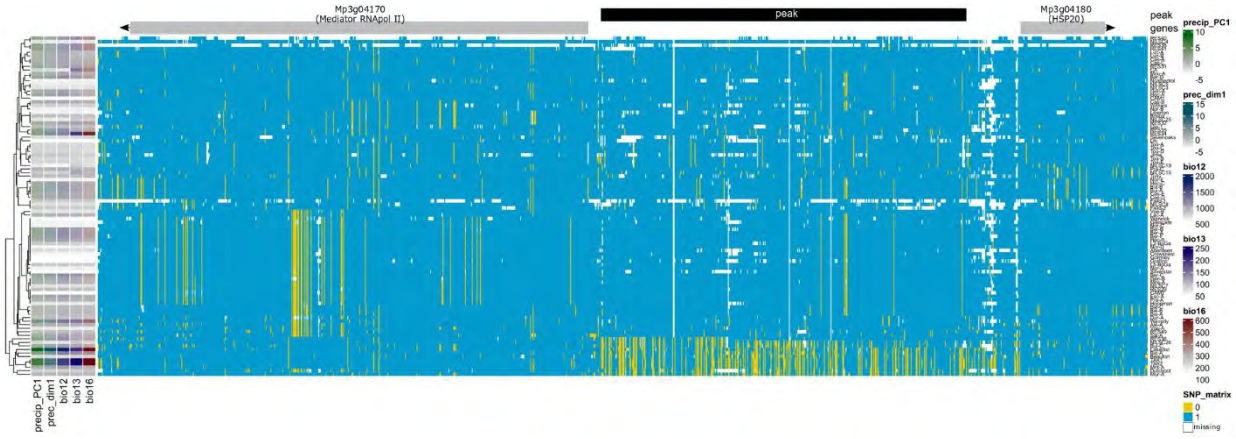


Figure 34: Heatmap representing the SNP matrix in the region of the GEA peak surrounded by the RNA II polymerase mediator and the HSP20. The accessions are clustered according to SNP allele similarity, and annotated with the climatic conditions with which the peak is associated.

HSP20 (or small heat shock proteins) have been shown to be involved in abiotic stress responses in angiosperms, making them a more rational candidate. These chaperones prevent protein aggregation in response to heat shock stress, but also to other abiotic stresses (and even biotic stresses). For instance, they show enhanced expression in rice during drought and salt stress, leading to enhanced tolerance **(Zou et al., 2012)**. Similar observations have been in other organisms (*E. coli* and yeast) expressing a rice HSP20 **(Guo et al., 2020)**, showing the ubiquity

of these mechanisms. Other HSP20 can enhance plant sensitivity to abiotic stresses, like an HSP20 from creeping bentgrass that negatively regulates plant response to environmental stress (X. Sun et al., 2016).

It has been suggested that specific HSP can be responsible for response to certain type of abiotic stresses. In an RNAseq analysis in response to various abiotic stresses (Tan et al., 2023), this gene is up regulated with heat stress alone, or crossed with other stresses (nitrogen deficiency, mannitol, salt, dark) and with combined nitrogen deficiency and light stress. These results, and the association of this gene with precipitation variable, suggests that this HSP could be involved in various stress responses like heat but also possibly water excess.

The lateral organ boundary transcription factor and unannotated candidate

A region of the chromosome 7 is associated with mean annual precipitations. The upstream gene is MpASLBD5 (Mp7g00650), and the downstream gene is not annotated (Mp7g00660). The peak is closer to the unannotated gene (Figure 35), but at its tail whereas it is potentially in the regulatory region of the MpASLBD5. It is therefore hard to determine which gene could be the causal one.

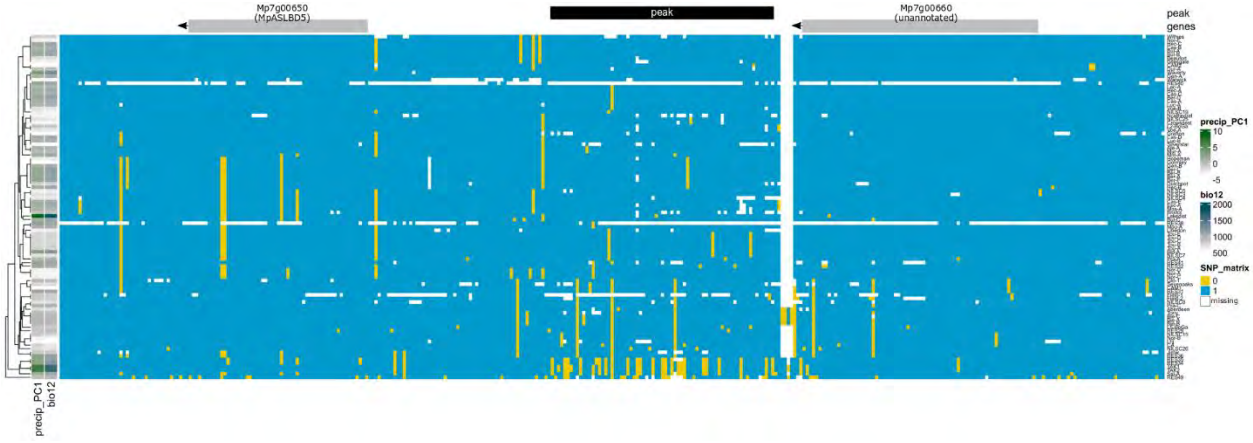


Figure 35: Heatmap representing the SNP matrix in the region of the GEA peak surrounded by the LOB domain protein and the unannotated protein. The accessions are clustered according to SNP allele similarity, and annotated with the climatic conditions with which the peak is associated.

In term of the genes functions, the only information associated with the unannotated gene is a PANTHER family id: cold regulated protein (PTHR34673). The candidate is a very short protein (74 AA) that is orthologous to the AT3G03341 gene in *A. thaliana,* of which little is known. In *M. polymorpha*, this gene is duplicated, which is not the case in angiosperms, and it is differentially expressed in response to many abiotic stresses: cold, light, heat, dark and many

combined stress conditions like mannitol+light, mannitol+N deficiency or salt+mannitol. This gene therefore seems to have a link with response to abiotic stresses, in *M. polymorpha* and perhaps in flowering plants, but its function has to be investigated.

The MpASLBD5 gene encodes for a plant specific transcription factor with a lateral organ boundaries domain. This type of genes was initially recognised as key regulator of plant organ development, but their role was then expanded to a variety of functions like plant regeneration, pathogen response or response to abiotic stress (Grimplet et al., 2017; C. Xu et al., 2016).

Both genes close to the peak are potentially linked with environmental conditions but need a bit more investigation to confirm their role.

The ribosomal protein, dynamin and peroxidase candidates

The chromosome 5 has a significant region linked with temperature and solar radiation variables (PC1 temp, PC1 all var, bio2 bio5 and PC1 srad). This region is overlapping a ribosomal protein L24e (Mp5g17510), and has a peroxidase upstream: POD128 (Mp5g17500), and a dynamin downstream (Mp5g17520). None of the three genes are differentially regulated under abiotic stress conditions from (Tan et al., 2023) study, and the ribosomal protein is under balancing selection, whereas the peroxidase is under selective sweep. The alternative allele is present in a non-negligible part of the collection (around 35 accessions) and correlated with high values of temperature and solar radiation (Figure 36).

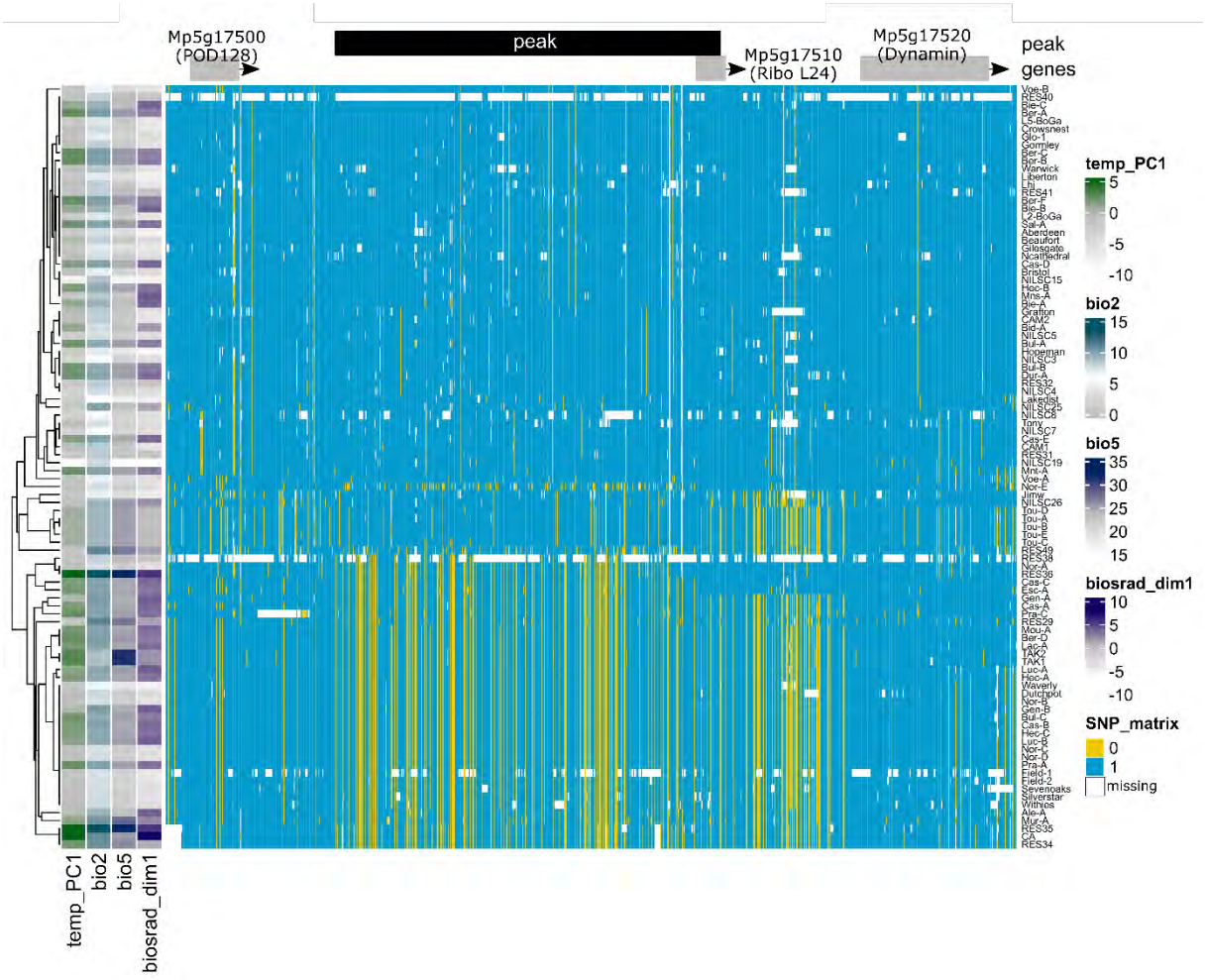


Figure 36: Heatmap representing the SNP matrix in the region of the GEA peak surrounded by the peroxidase, ribosomal protein and dynamin. The accessions are clustered according to SNP allele similarity, and annotated with the climatic conditions with which the peak is associated.

Regarding the function of the surrounding genes, L24e ribosomal proteins are found in eukaryotes and archaea and play a role in the kinetics of peptide synthesis. Dynamins are multidomain large GTPases, involved in endocytosis and intracellular trafficking, playing a role in a variety of cellular processes. This candidate gene is orthologous to a dynamin related protein 5A (AT1G53140) of *A.thaliana,* that is involved in cytokenesis. The peroxidase belongs to a group of 25 *M. polymorpha* class III peroxidases, five of which are located in a gene cluster with the candidate (Mp5g17450, Mp5g17460, Mp5g17470, Mp5g17480, Mp5g17490) **(Passardi et al., 2007)**. This group originates from a *M. polymorpha* specific duplication, since there is only 2 orthologs outside of *M. polymorpha*, both in *M. paleacea* (web page of the HOG concerned: https://peroxibase.toulouse.inra.fr/orthogroups/view_orthogroup/Prx049). Peroxidases are involved in development (control of auxin levels and hydroxyl radical generation thereby influencing cell elongation…) as well as in adaptation with resistance to

dehydration, cold or pathogens via their role in lignification, oxidative bursts, antioxidant protection or activation of the stress response signalling cascade (Kidwai et al., 2020). This gene may therefore be linked to *M. polymorpha's* response to higher temperature.

The ERECTA candidate

A region of chromosome 4 is associated with precipitations and solar radiation (PC1 precip, PC3 all var, PC1 month precip, bio12 bio13 bio16 bio18, PC2 srad). Upstream is lying an ankyrin repeat gene (Mp4g14600) and downstream is an unannotated gene (Mp4g14610) that overlaps with an ERECTA gene MpER (Mp4g14620). The ankyrin repeat is close to the AKR and EMB506 genes of *A. thaliana* (AT5G66055, AT5G40160), involved together in plastid differentiation, morphogenesis, and organogenesis, and the mutant is embryo defective. If the *M. polymorpha* gene has conserved this tightly controlled function, it is an unlikely candidate of environmental adaptation. The ERECTA gene is 4 kb from the peak thus the latter could impact its regulatory region. This gene is the pro-ortholog of the three *A. thaliana* genes ERECTA, ERECTA like 1 and ERECTA like 2. This receptor like kinases gene is known for its role in the plant development but also in modulation of signalling pathways in response to the environment (Van Zanten et al., 2009), like the control of stomatal transpiration efficiency in *A. thaliana* (Masle et al., 2005).

The LOX and Cupin candidates

A region on the chromosome 1 is overlapping Mp1g21940 and is surrounded by the genes MpLOX5 (Mp1g21930) and MpCupin4 (Mp1g21950). The overlapping gene has a DNA-binding domain and seems to be the ortholog of the CLPF gene in *A. thaliana* (AT2G03390) that participates in protein homeostasis in the chloroplast (Nishimura et al., 2015). The gene upstream of the peak is a lipoxygenase, an oxidoreductase enzyme that acts on lipids and lead to the production of various oxylipins, including jasmonic acid (JA). Our candidate is closest to the LOX2, 3, 4 and 6 gene (AT3G45140, AT1G17420, AT1G72520 and AT1G67560 in *A. thaliana*), that are indeed involved in production of JA precursors (T.-H. Yang et al., 2020). Since JA is linked with plant response to biotic and abiotic stresses, this LOX could be involved in *M. polymorpha* adaptation to climatic conditions. The gene downstream is a cupin, a very diverse protein family that is involved, among other things, in development, stress response or defence against pathogens (F. Hu et al., 2023). The candidate gene does not seem to have a close ortholog in angiosperms, and was only found in liverworts, mosses, hornworts, gymnosperms,

and charophyte with a e-value threshold of 10^{-5} (for angiosperms, only a hit was found in *Cucumis sativa,* and is placed in the hornwort clade, suggesting an error). This means that this type of cupin was present in the common ancestor of land plants and has been lost in angiosperms. It is therefore hard to hypothesize on its function regarding Marchantia adaptation.

3) Concluding remark on the GEA

The detail of the candidates harboured in the regions highlighted by the GEA analysis gives a first idea of what may be the important genetic factors of *M. polymorpha's* adaptation to different climatic conditions, and perhaps a first insight into bryophyte's genetic adaptation to climate. It is not always possible to determine the most promising genomic regions and the true causal candidates in the zones only based on bioinformatics analysis. Since Marchantia is a quite recent model, for most of the genes found, it is only possible to speculate on the role that they may have in *M. polymorpha's* adaptation, based on imperfect comparisons like the ones with the closest orthologs from the well-studied model *A. thaliana*. But still, some promising candidates were found, some already known in flowering plants, like the ABC1K7, and some that raises more Marchantia-specific questions like the MpIDS1, with the link between oil bodies and abiotic stresses. These candidates represent a great resource for functional validation in the wet lab.

Interestingly, the accessions bearing the alternative haplotype generating the association signal are often the same ones, for the different candidate regions. On the Figure 37, we can see that accessions that often bear the minor haplotype are from areas with values of temperature, precipitations or both that differ from the majority of the Marchantia collection (like TAK1, CA or Lakedist). Quite strangely, the NILSC accessions that come from Scandinavia do not seem to intervene in the genomic signals detected despite their extreme life conditions. The reference genome for *M. polymorpha* is from TAK1, that originates from a location with climatic conditions quite different from the rest of the collection and often bears the alternative allele. This sometimes altered the quality of the mapping of other accessions on its genome, as it can be seen with the missing SNP data on Figures 29 & 33 (ABC1K and NLR candidates). Using

another reference genome to determine the markers used in this GEA could therefore refine the results and allow to map more precisely the causal loci.

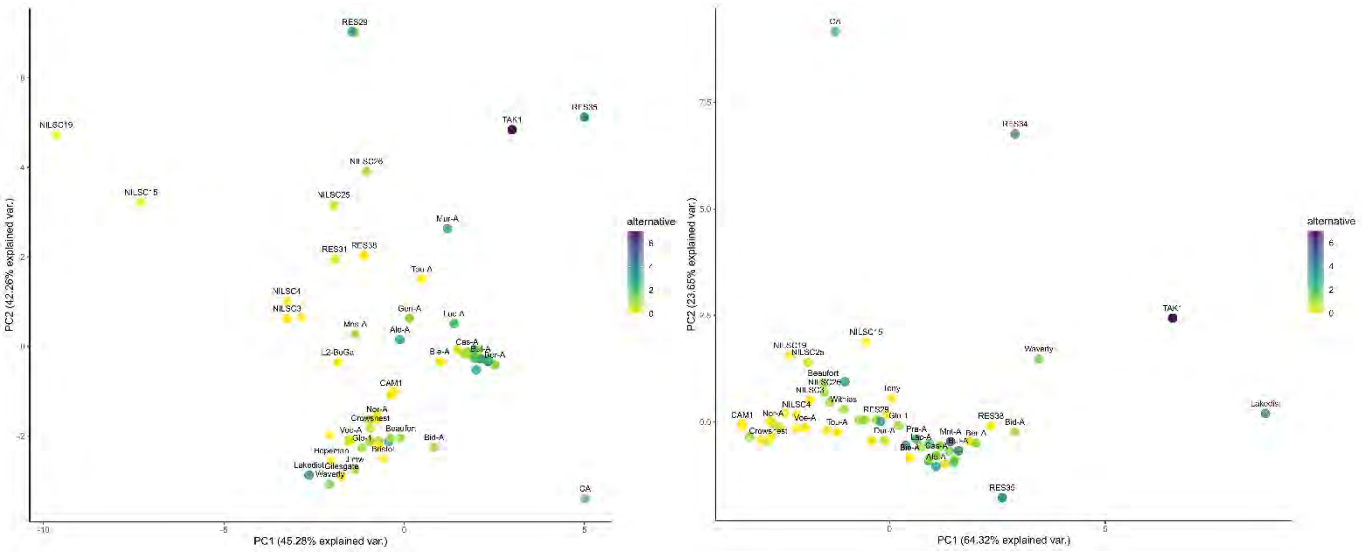


Figure 37: Position of the accessions relative to the PC1 and PC2 of the PCA on temperature (left) and precipitation (right) bioclimatic variables.

The accessions are coloured depending on the number of times they bear the alternative haplotype in each of the 8 candidate regions mentioned above. A darker colour accession bears the minor haplotype in more candidate regions than a lighter colour one.

The power of this GEA would be greatly enhanced with a broader and more homogenous sampling of plants, but also with measures performed in the microenvironment where the plants were found. Indeed, Marchantia are often found sheltered, close to water points, and the broad climatic measures from the Worldclim data can be quite different from the real-life conditions in the microenvironment of these plants. Another concern about the nature of our data could be the exact extent to which *M. polymorpha* accessions are actually mobile. Indeed, it seems that there is significant gene flow between *Marchantia polymorpha* individuals, almost at a worldwide level (cf chapter 1), and this gene flow could be partly due to the human-driven transport of these plants. To run a GEA the organisms studied should have adapted to their environment, and this adaptation should have left a mark in their genome. But if we hypothesise that a non-negligible number of *M. polymorpha* have only recently arrived at the location they were sampled, the power of the GEA would be reduced.

II) GWAS of the response of *M. polymorpha* to a fungal pathogen

The genome-environment association study allowed identifying some candidate genes for the adaptation of *M. polymorpha* to abiotic stresses. To study the response of the bryophyte to biotic stress, a genome wide association study was carried out on the evaluation of *M. polymorpha's* response to *Colletotrichum nymphaeae*, an hemibiotrophic fungal pathogen. The strain used here was isolated on *M. polymorpha* by Jessica Nelson **(J. M. Nelson et al., 2018)** as a *Colletotrichum* sp 1 strain, later characterised by our team as a *C. nymphaeae* strain.

1) Material and methods

a) Experimental design and data modelling

Different accessions of *M. polymorpha* (87, among which 77 belong to the *ruderalis* subspecies) were inoculated with *C. nymphaeae* by Karima El Mahboubi (PhD student in the LRSV's EVO team). The plants were then phenotyped by quantifying the area of their thallus or the browning area caused by the fungal infection. The area of the thallus was measured pre-inoculation and 6 days post inoculation (6dpi), and the browning area was measured at 6 days post inoculation. The measure time point of 6dpi was chosen because after that time, the symptoms barely evolve in most accessions. Different phenotypic variables were calculated based on these observations and are detailed in Table 3.

Table 3: Names of the phenotypic variable used in the GWAS analysis and detail of how they were obtained

Variable name	Explanation
Thallus_area_6dpi	Measure of the thallus area at 6 dpi (developmental phenotype)
Brown_area_6dpi	Measure of the browning area on the thallus at 6 dpi
Brown_perc	Brown_area_6dpi/thallus_area_6dpi*100 (proportion of the thallus infected at 6 dpi)
Preinoc_thallus_area	Thallus area pre-inoculation
Diff_area_6dpi_ni_cnm	Difference of thallus area between inoculated and non-inoculated plants at 6dpi

These variables allow to study the response of *M. polymorpha* to *C. nymphaeae's* infection, but also to start delving into developmental questions by studying the non-inoculated plants (the negative controls) and their thallus area.

The inoculation and phenotyping of the different accessions were performed according to an experimental plan that would allow the results to be homogenous. The phenotyping of 87 accessions could not be executed all at once, therefore nine batches of accessions were processed, each time with two internal controls present in all batches: TAK1 (the reference accession) and CA (an accessions showing resistance compared to TAK1). All accessions have been processed in at least 1 experimental batch, with two different phytotrons containing each one inoculated and one non-inoculated petri dish of the accession, piled up together (Figure 38). Since the petri dishes contain 9 plants, at least 18 plants were phenotyped for each accession and each condition (inoculated or non-inoculated with the pathogen). The inoculation with the pathogen (or a mock solution for the negative controls) was performed 15 days after the transplanting of the gemmae.

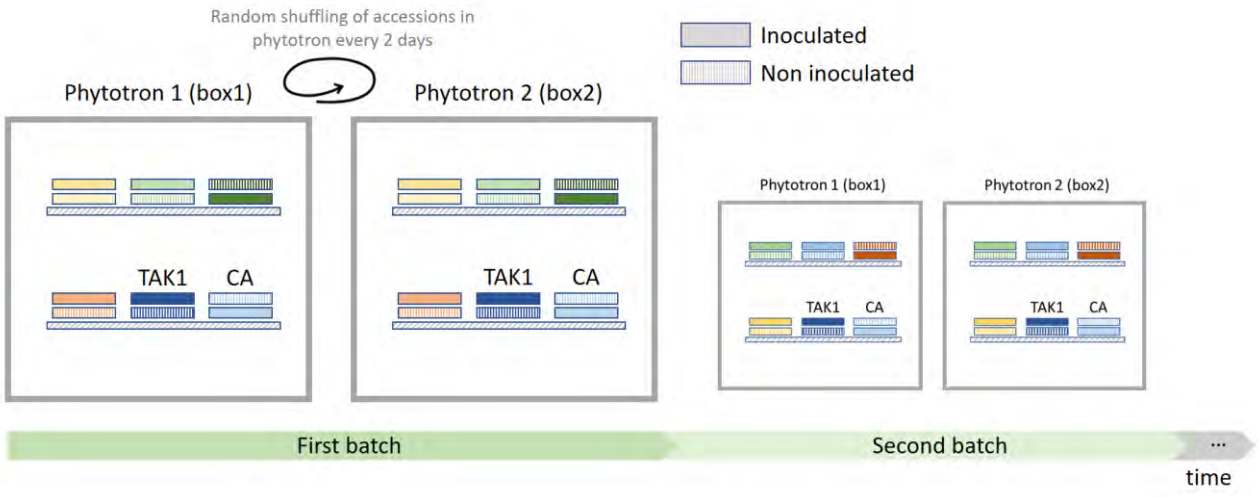


Figure 38: Experimental design used for the phenotyping of the 87 accessions of M. polymorpha inoculated with C. nymphaeae. Multiple batches were processed along time to phenotype all the different accessions. In each batch, the plants of each accession were separated between two phytotrons with an identical layout: for each accession the inoculated and non-inoculated petri dishes are piled up, the location of an accession is identical in both phytotrons and is changed randomly every two days (including change of the upper petri dish in the pile of each accession). In each batch the two controls (TAK1 and CA) can be found among the accessions.

For the GWAS analysis, a mean value of each phenotype of interest is needed. Given the experimental plan, several elements can bias these mean values:

- disparities between the two phytotrons

- disparities between the different batches

- the thallus area of each plant before inoculation

The raw phenotypic data were therefore corrected before their use as inputs of the GWAS. First, the plant being outliers (with their value being higher or lower than 1.5 times the interquartile range) for at least two phenotypic variables were discarded. Then, a linear model was applied on the data for each phenotypic variable (except the thallus area pre-inoculation, that is used as a covariable), distinguishing inoculated and non-inoculated conditions when needed. The phenotypic variables separately corrected are therefore: thallus_area_6dpi inoculated, thallus_area_6dpi non inoculated, brown area inoculated and brown_perc inoculated.

In the case of TAK1 and CA, that are present in each batch, the linear model takes into account the effect of each experimental batch, the effect of the two phytotrons and the effect of the thallus area before inoculation. The interactions between the phytotron and the experimental batch are also taken into account, leading to this model:

```
Phenotype (TAK1 or CA) ~ phytotron * batch + preinoc_thallus_area
```

This allows to get a global corrected mean for CA and for TAK1 (used as their phenotypic value in the GWAS).

Another linear model is implemented in order to use TAK1 as an internal control of the experimental batches:

```
Phenotype (TAK1) ~ batch + preinoc_thallus_area
```

This model only takes into account the experimental batch and the area of TAK1 plants before inoculation, because it will be used as a covariable for the other accessions. Each phenotype measured on an accession is associated with the internal control TAK1 in the same experimental batch, and corrected according to the following linear model:

```
Phenotype ~ acc * phytotron + preinoc_thallus_area + TAK1_ctrl
```

acc being the effect on the phenotype considering a given accession, and TAK1_ctrl TAK1's corrected mean under the same experimental batch.

For each phenotype, estimated marginal means of each accession are calculated with this model, and serve as inputs in the GWAS. These means are represented on the Figure 39 (diamonds), compared to the raw data (boxplots), for the thallus area at 6dpi in non-inoculated plants, and for the brown area at 6dpi for inoculated plants.

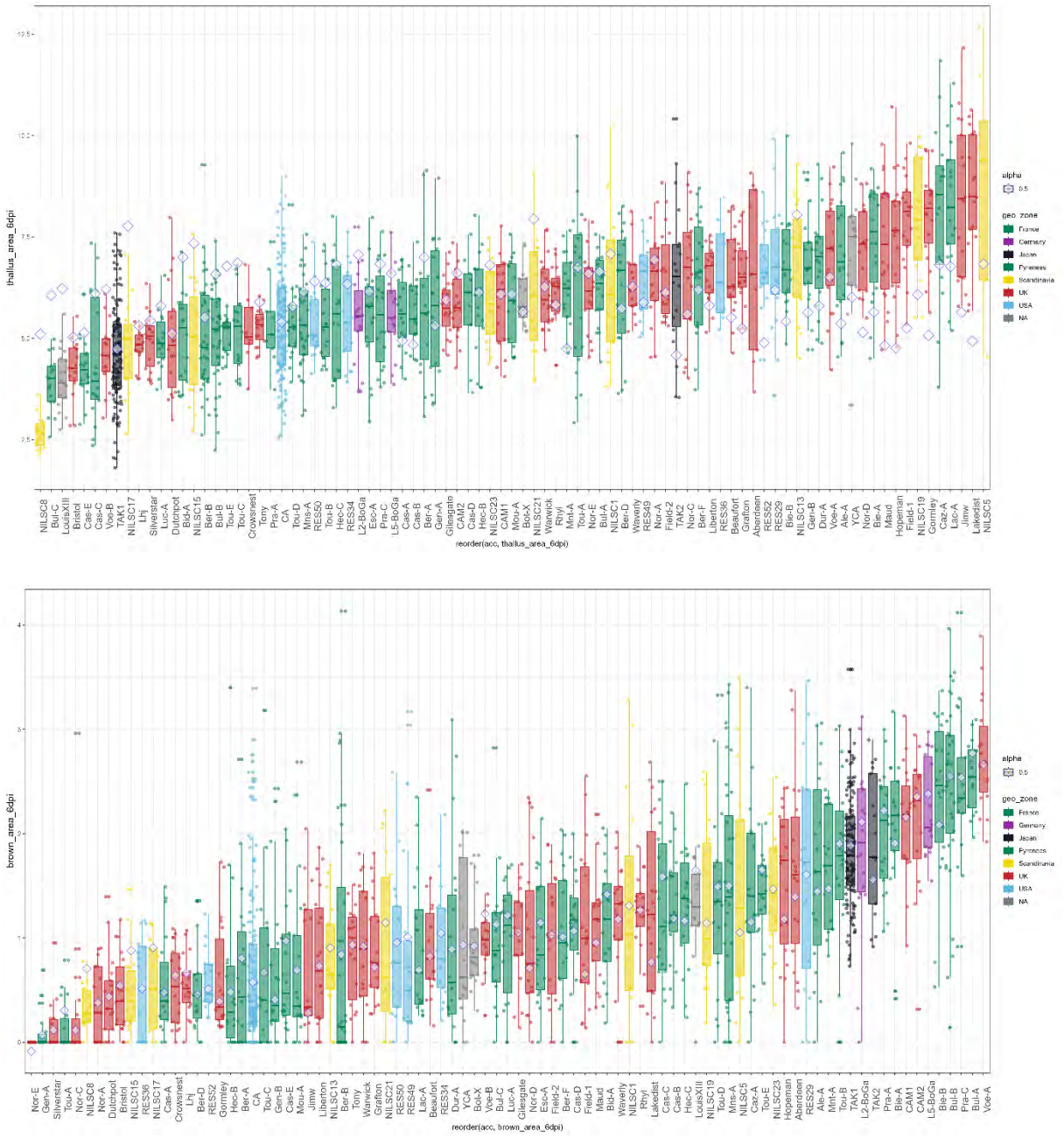


Figure 39: Variation of the phenotypes from the 87 M. polymorpha accessions: thallus area of non-inoculated plants on the upper part, and brown area of inoculated plants on the lower part.The raw measures are represented by the boxplots and the corrected means are symbolised by the diamonds.

The size of the thallus and the susceptibility of *M. polymorpha* to *C. nymphaeae* display variation across the phenotyped accessions. This variation does not seem to be correlated to

the geographic origin of the accessions. For the brown area measure, the corrected means does not change a lot the relative order of the accessions compared to the raw data. For the thallus area at 6dpi it changes it a bit more, probably due to the effect of the correction with the thallus area pre-inoculation, that is correlated with the area at 6dpi.

Based on these corrected phenotypes, a last variable to be tested in GWAS was calculated: Diff_area_6dpi_ni_cnm, which is the difference of the thallus area at 6dpi between non inoculated and inoculated plants from the same accession. This allows to evaluate the developmental repression associated with the infection.

b) GWAS model implemented

Among the 87 accessions phenotyped, only the 77 belonging to the subspecies *ruderalis* will be used for the genome wide association study, to avoid the confounding effect of the strong genetic structure between the three subspecies. The SNP data for the 77 accessions was therefore extracted from the *ssp ruderalis* VCF, with no MAF or missing data filter, in order to be able to perform the independent association analysis on all sites and then to test differential filters of minor allele frequency MAF and missing data, leading to approximately 4.8 M of markers.

The GWAS analysis was performed in the same fashion as the one of the GEA. The GEMMA software was used, with the linear mixed model corrected with a centered relatedness matrix, and each phenotype was tested independently in the univariate linear mixed model with the options `–lmm 4 –maf 0 –miss 1`, leading to association results for each variant. The resulting p-values are displayed in the qqplots of Figure 40 (for these qqplots, SNPs were filtered for a MAF of 5%, maximum 15 individuals with missing data at one site and had to be on one the 8 autosomes).

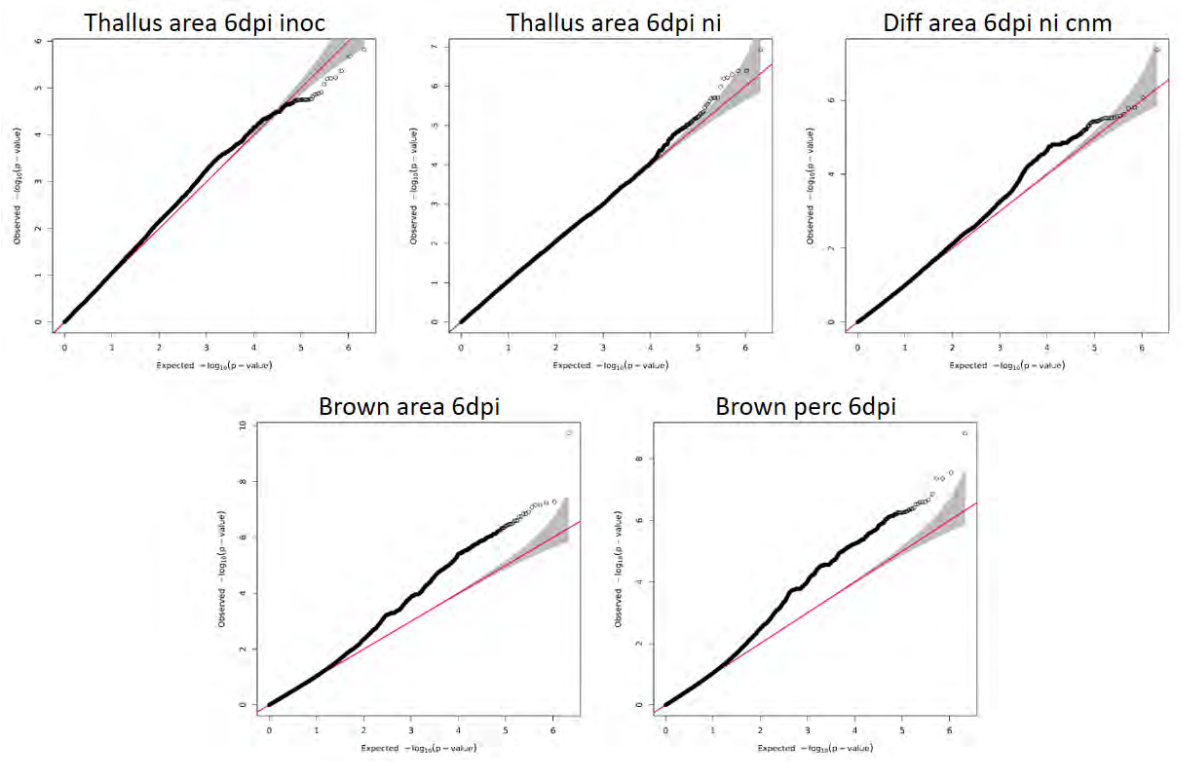


Figure 40: QQplots of the p-values obtained for the 5 phenotypes evaluated in the GWAS. The X axis is the -log(p-value) for p-values that follow a uniform distribution (expected p-values), whereas the Y axis corresponds to the -log(p-value) of p-values obtained when evaluating the association of SNPs with the phenotype. The red line corresponds to a 1 for 1 relationship between the X and Y axis. The grey area represents the 95% confidence interval for the QQplot under the null hypothesis of no association between the SNPs and the phenotype. First line of plots are phenotypes linked to thallus area and second line are browning phenotypes in response to C. nymphaeae.

The QQplots of the brown area related phenotypes show SNPs departing from the uniform distribution for quite low -log(p-value), but the local score approach will allow to pinpoint only the most significant regions with cumulated association signal. The local score was therefore used on these results, with a ξ value of 2 to determine SNPs contributing to the Lindley process (only SNPs with a p-value smaller than 10^{-2} contribute to the local score process), and chromosome-specific significance thresholds based on a resampling strategy. All the local score peaks higher than this significance threshold were extracted, and they were annotated with their overlapping (when existing), downstream and upstream genes. All the resulting candidates are available in SupData2.3.

2) Candidate regions

a) Description of the different candidate regions for the different phenotypes

Developmental phenotypes

First the "developmental" phenotypes were scrutinised. The thallus area at 6dpi in inoculated plants and in non-inoculated plants both have five significantly associated genomic regions, three of which are common between the two phenotypes (Figure 41). These three regions should be linked to processes influencing the thallus size, regardless of the infection status of the plant. One of the shared regions (on chromosome 4) is surrounded by GSDL lipases (Mp4g07030, Mp4g07040), a very diversified family of serine hydrolases (around 100 members in flowering plants), some of which are involved in cutin synthase **(B. Xu et al., 2021)**. This type of protein participates in plant growth and development, but also in plant response to abiotic stresses, and the closest *A. thaliana* gene to our candidates are guard cell enriched GDSL lipases (GGL29 AT5G62930, GGL28 AT5G45920…), highly expressed near the stomata **(Xiao et al., 2021)**. This type of protein is part of *M. polymorpha* cell wall proteome, and could therefore be involved in thallus growth **(Kolkas et al., 2022)**. The other peak (also on chromosome 4) is surrounded by two UDP glucosyl transferases (MpUGT18 (Mp4g16850) and MpUGT19 (Mp4g16860)), that are similar to *A. thaliana* coniferyl alcohol glucosyltransferase (UGT72E1,2,3) and therefore could be involved in phenylpropanoid glucosylation **(Lanot et al., 2006)**, influencing the pre-lignin pathway in the cell walls and cuticle of *Marchantia* **(Espiñeira et al., 2011; Renault et al., 2017)** and have an impact on its thallus development. This exact same region was associated with annual precipitation (bio12) in the GEA, but does not appear to be differentially regulated in stress conditions, except for the third day of infection by *P. palmivora* **(Carella et al., 2019; Tan et al., 2023)**.

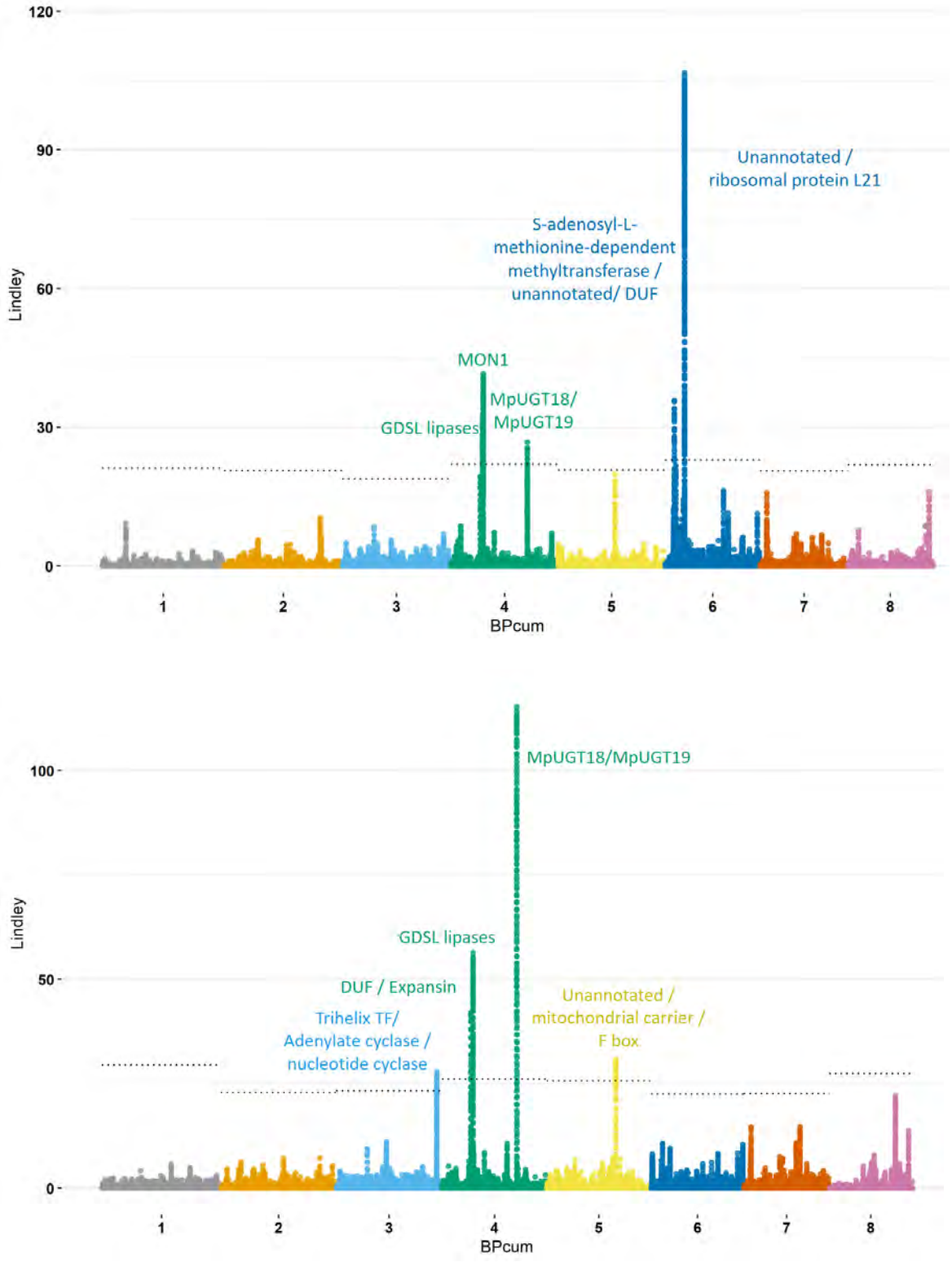


Figure 41: Manhattan plots of the GWAS on the thallus area of the non-inoculated (upper plot) and inoculated (lower plot) plants at 6dpi. The peaks resulting from the local score processing of the data were annotated with the surrounding genes functions.

Then some regions were only detected in association with non-inoculated plants thallus area. There is one on chromosome 4, closely following the GDSL peak, that is surrounded by a non-

annotated gene (Mp4g07090) located 39 kb upstream of the peak and a Monensin sensitivity 1 (MON1) gene (Mp4g07110) closer to the peak (2 kb away). The MON1 gene may contribute to the plant growth through its role in vacuolar trafficking **(M. K. Singh et al., 2014)**.

Two other candidate regions are on the chromosome 6, the first one overlapping with an unannotated gene (Mp6g01790) under balancing selection and the second one is upstream of a ribosomal protein L21 (Mp6g04500), probably in its regulatory region. Differences in ribosome operation could result in differences in plant growth.

For the inoculated plants, a specific region associated with thallus area is the one on the third chromosome, potentially in the regulatory region of a trihelix transcription factor (TTF, MpTRIHELIX36 Mp3g24320) that displays a really poor mapping for the accessions with the bigger thallus (Figure 42), suggesting a presence/absence polymorphism of this genomic region, which could have an effect on thallus area. TTF are known to be involved in plant development, response to biotic and abiotic stress and phytohormone response **(X. Cheng et al., 2019)**.

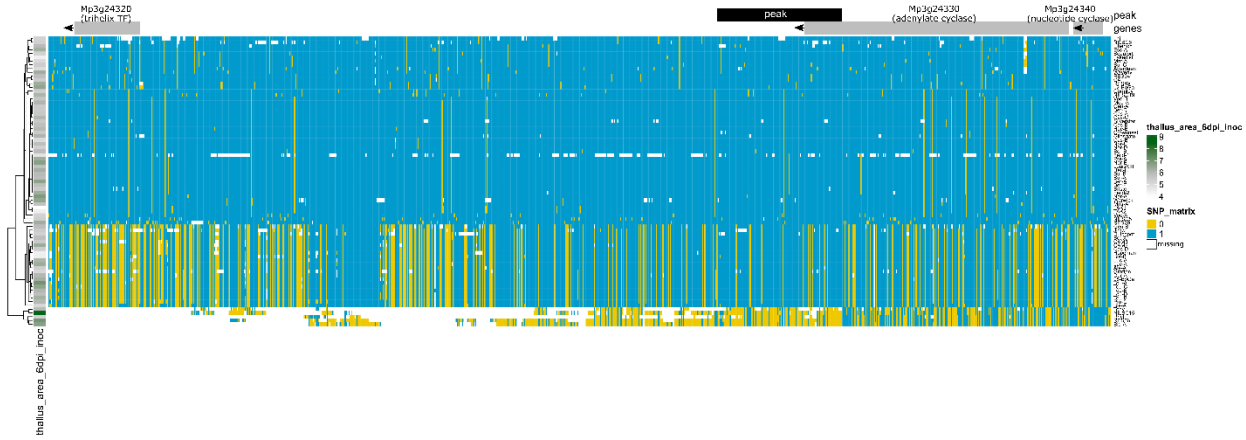


Figure 42: Detail of the alleles on the SNPs located close to the peak of the trihelix transcription factor (chromosome 3). The accessions with the bigger thallus area have missing data in the neighbourhood of the trihelix transcription factor, suggesting the absence of this region in these accessions.

The second region associated with the thallus area of inoculated plants is on the chromosome 4 that is closely followed by an expansin (Mp4g06560). Expansins participate in cell wall plasticity, allowing the cells to expand and to rearrange, being therefore crucial to plant growth and development **(Marowa et al., 2016)**. This candidate has enriched expression in air pore cells **(Wang et al., 2023)**, and is down regulated during infection by *P. palmivora* and different

abiotic stresses (N deficiency, dark, heat and combination of stresses…) (Carella et al., 2019; Tan et al., 2023).

The final region is 12 kb away from an F-box protein (Mp5g16770). This protein family, one of the largest in plants, mediates protein degradation through the formation of an E3 ubiquitin ligase complex. They are therefore involved in a plethora of biological processes, among which is plant growth and development (Abd-Hamid et al., 2020). This one is up regulated during infection with *P. palmivora*, and with some abiotic stresses (salt, cold and nitrogen deficiency combined stress), and is under balancing selection in the population (top 1% of genes under balancing selection, with high Tajima's *D* values).

The results of the genome wide association with the difference between thallus area in inoculated and non-inoculated plants from the same accession were also inspected, to try to understand how the infection can affect the plant growth (Figure 43). The two peaks located on the chromosome 8 are hardly exploitable because one is only surrounded by unannotated genes, and the other one spans over a region containing a lot of different genes, making it difficult to pinpoint a causal candidate. The two remaining peaks are located on chromosome 4 and 7. The peak on chromosome 4 overlaps an RNA binding protein (Mp4g11580) potentially involved in pre-mRNA splicing (similar to SR45 - AT1G16610 - protein from *A. thaliana*) and is potentially located in the regulatory region of a nuclear transport factor 2 protein (NFT2, Mp4g11570). NFT2 plays a role in the import of proteins in the nucleus (Q. Zhao et al., 2006), and is down regulated in *M. polymorpha* in response to a lot of stresses, including infection by *P. palmivora* (4dpi). The peak of chromosome 7 is preceded by a glycerol 3-phosphate acyl transferase (GPAT6, Mp7g00920), that synthesises the precursor of cutine, a physical barrier on the epidermis of the plant (Kong et al., 2020), and whose expression is enriched in cells surrounding the notch and air pore cells (Wang et al., 2023). This candidate's expression is down regulated during a variety of stresses, including *P. palmivora* infection (1 and 3 dpi). It is followed by a protein from the 2-oxoglutarate and FeII dependant oxygenase superfamily (Mp7g00930), that is similar to cupuliformis2 (AT3G18210) and incurvata11 (AT1G22950) in *A. thaliana*. These epigenetic repressors act in concert to regulate the organ development and life cycle of the plant (Mateo-Bonmatí et al., 2018). Both could be linked to the differential growth responses of accessions to the infection.

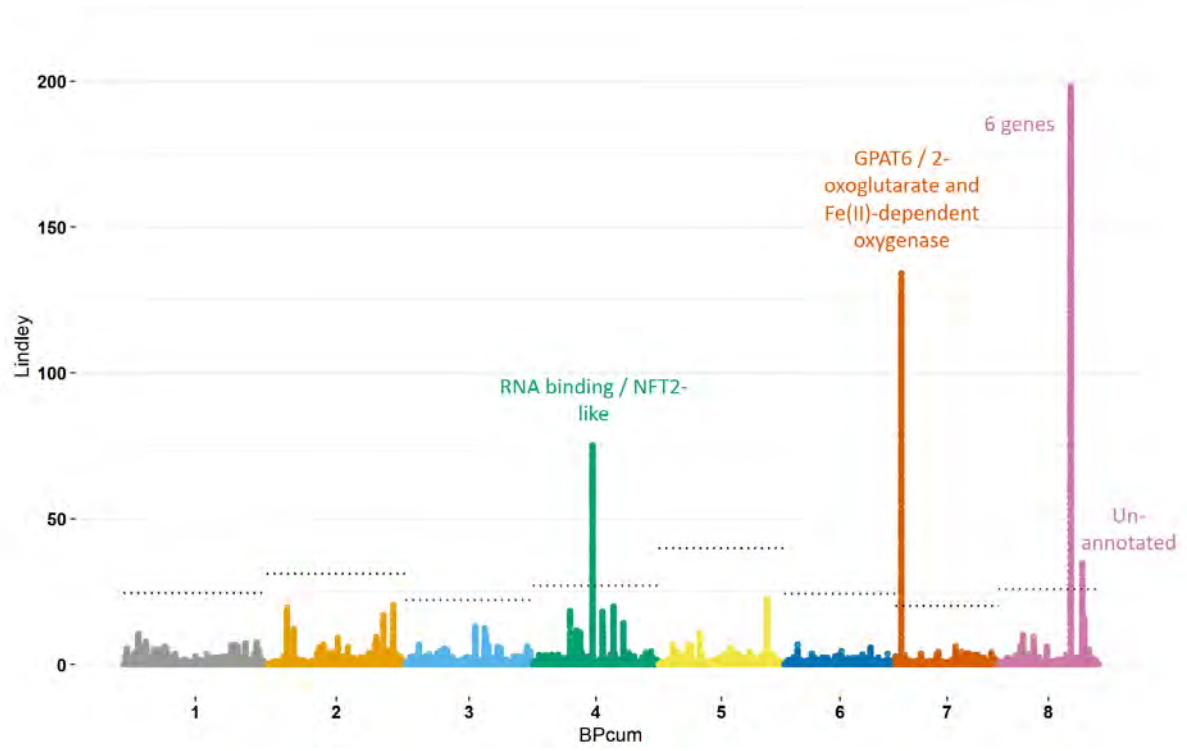


Figure 43: Manhattan plot of the association with the thallus area difference between infected and non-infected plants. The peaks resulting from the local score processing of the data were annotated with the surrounding genes functions.

A lot of the genomic regions found here harboured unannotated genes, or genes whose exact function is hard to determine. It is therefore complicated to conclude on the different parameters that determine the differences in thallus size among the accessions, except from the processes associated with cell walls plasticity. Indeed, many proteins involved in cell wall modification (Expansins, GDSL, UGT) were found in this GWAS analyses.

The thallus size and the susceptibility of the plants do not seem linked (Figure 44), but we tried to investigate the mechanisms impacting the thallus size during the infection. Indeed, the accessions display variable growth behaviour under infection: some display impaired growth and a smaller thallus size than the non-infected plants (like Voe-A), while some other seem to have bigger thalli under infection (like NILSC15). The latter could be a strategy to avoid the pathogen. The interpretation of most association signals was impaired by the presence of unannotated genes or the width of the signal, that makes it hard to determine the causal gene. But the only peak precisely annotated suggested a role of epigenetic mechanisms or of the cuticle in the variation of the pathogen's impact on thallus growth.

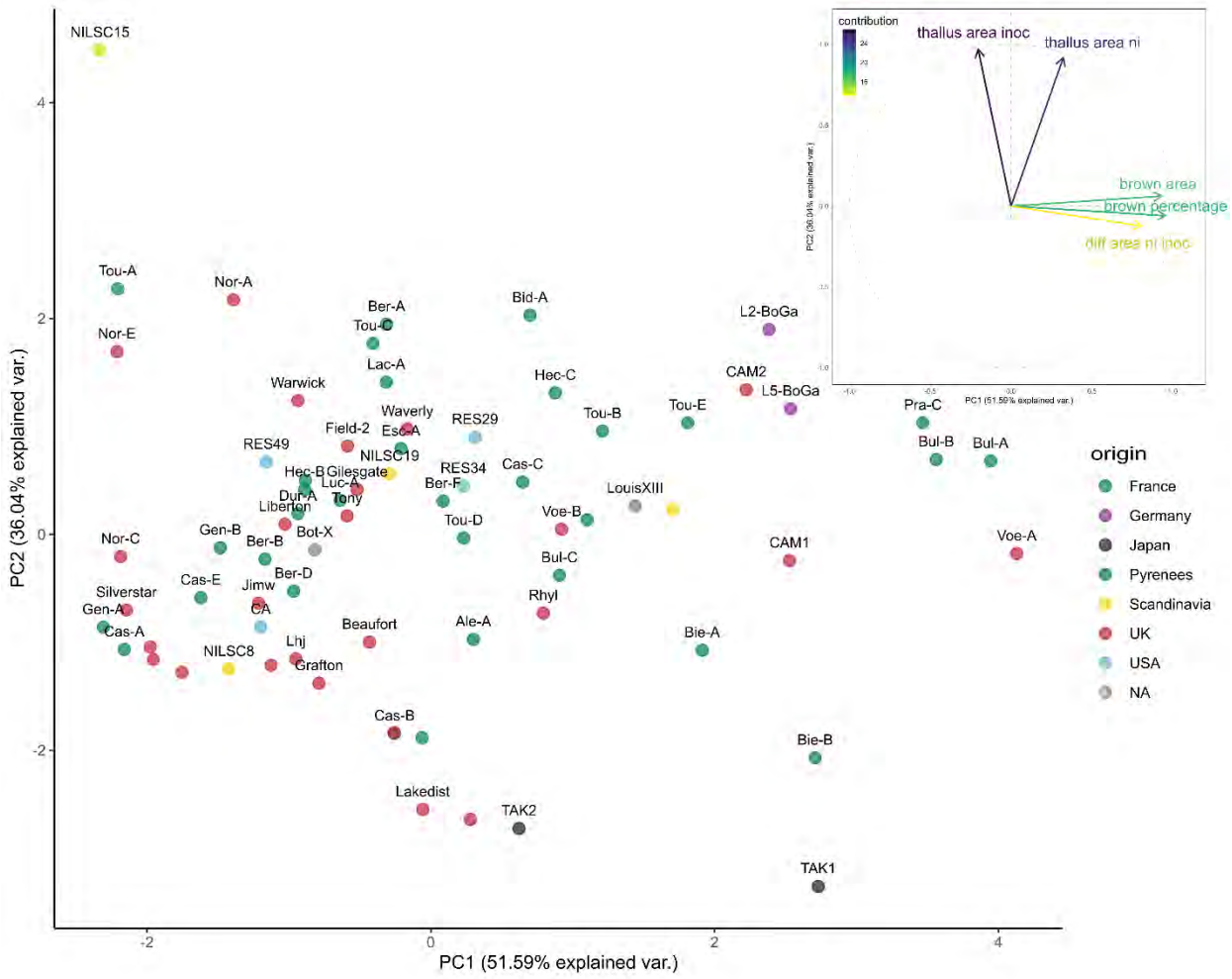


Figure 44: Accession position relative to the two first components of the PCA on the five phenotypes from the GWAS. The PC1 is linked to symptoms phenotypes and the PC2 is linked to thallus area. The response to the pathogen and the thallus area do not seem correlated. The geographic origin of the accessions does not seem to condition their phenotypes either.

Symptom-related phenotypes

The main objective of this GWAS is to uncover the mechanisms used by *M. polymorpha* to respond to *Colletotrichum nymphaeae*. The brown area on the thallus caused by the fungal infection was therefore studied, as well as a derived variable, being the proportion of the thallus covered by the browning (Figure 45). Five genomic regions were detected as significantly associated to each of these phenotypes, three of which are common between the two phenotypes.

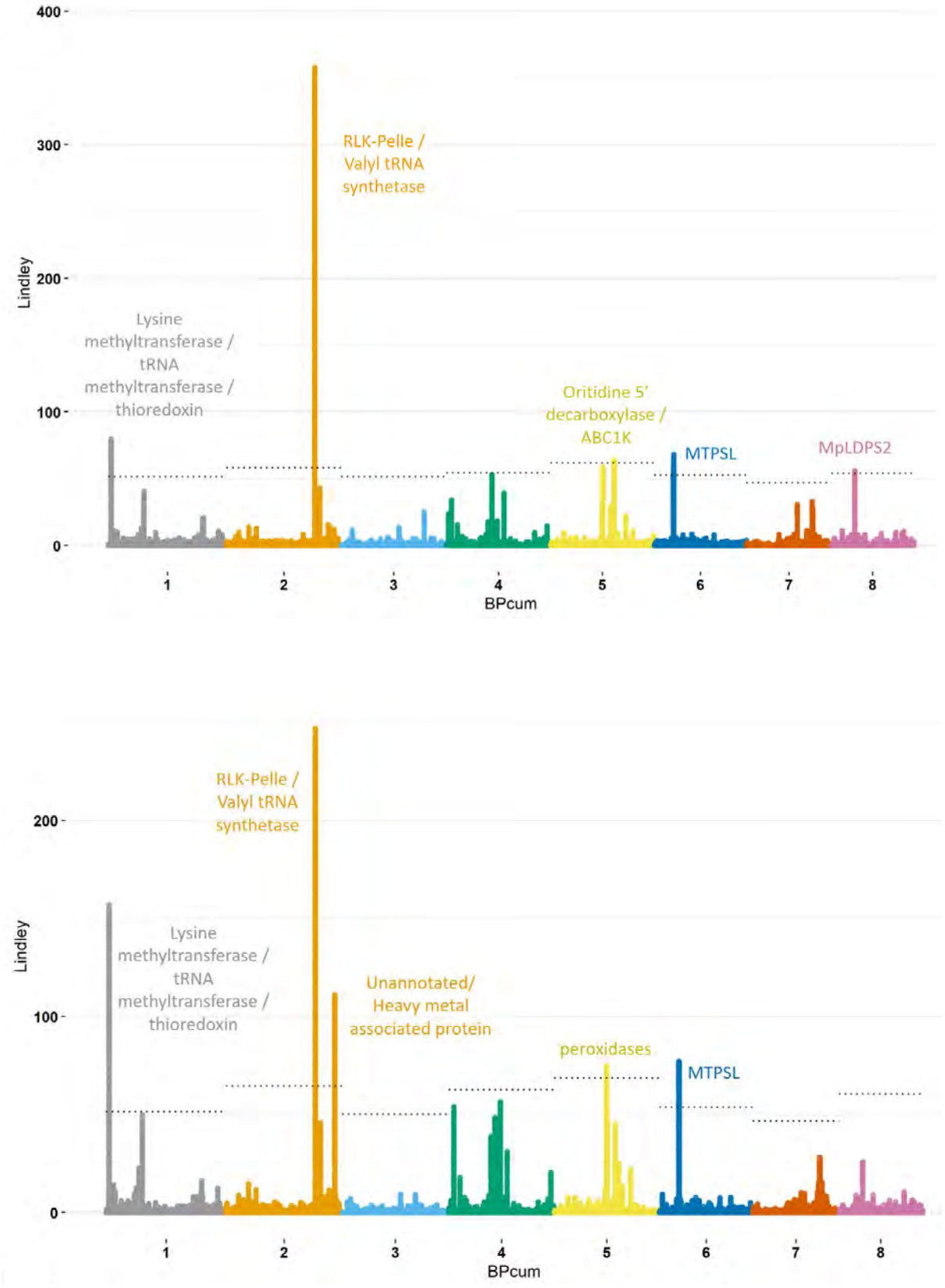


Figure 45: Manhattan plots of the GWAS on M. polymorpha's response to fungal infection at 6dpi. The GWAS on the brown area (upper plot) and on the percentage of the thallus covered in brown (lower plot) are represented after application of the local score and annotated with the surrounding genes functions.

The first peak shared by the two linked phenotypes is located on the chromosome 1, in the regulatory region of a thioredoxin (Mp1g00650). These proteins are involved in plant immunity by their reductase activity on cysteine residues that function as signalling switches, as well as their role in ROS detoxification **(Mata-Pérez & Spoel, 2019)**.

The second peak, on chromosome 2, is surrounded by a tRNA synthetase and a receptor-like kinase (Mp2g20720). Four of the most susceptible accessions share an alternative allele in this genomic region, potentially coupled with a small deletion, represented by the missing SNP values on Figure 46. This receptor-like kinase is in the same orthogroup as 15 *A. thaliana*'s brassinosteroid signalling kinases (BSK) (according to the orthogroup analysis carried for the interspecific comparison of selection signatures, cf chapter 1). In flowering plants, brassinosteroids have been shown to modulate innate immunity and orchestrate crosstalk between defence signalling pathways **(De Bruyne et al., 2014)**. Brassinosteroid synthesis being conserved in all land plants **(Yokota et al., 2017)**, this RLK could therefore be involved in brassinosteroid mediated immunity.

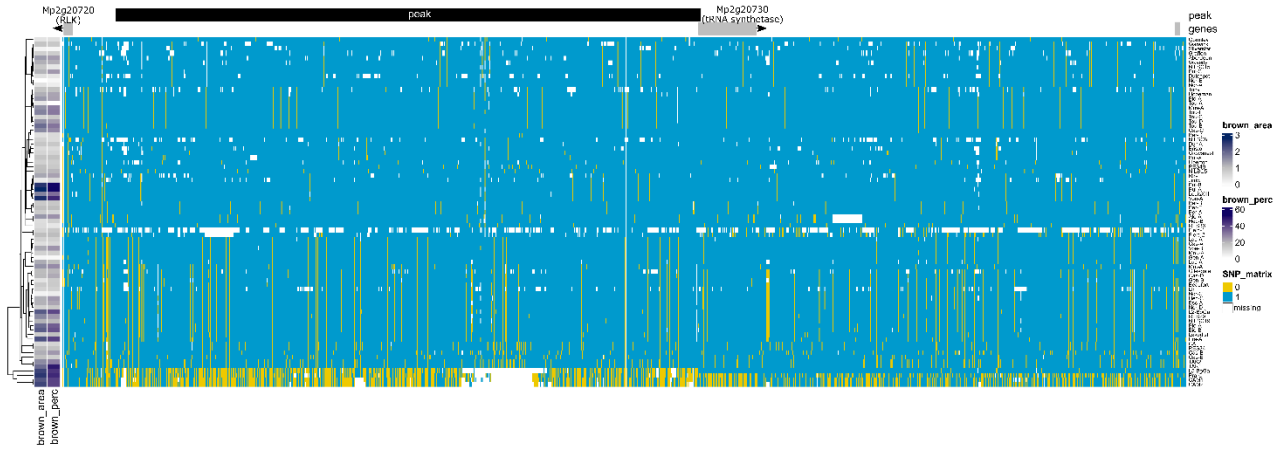


Figure 46: Detail of the alleles on the SNPs located close to the peak of the receptor like kinase (chromosome 2). Some of the accessions with the bigger browning area in response to the pathogen display a common haplotype (bottom of the SNP matrix), with a potential deletion.

The third peak is located on the chromosome 6 and overlaps with a cluster of 4 terpene synthases (Mp6g04580, Mp6g04590, Mp6g04605, Mp6g04610) and followed by a fifth one (Mp6g04630). These terpene synthases do not belong to the family classically found in plants (plant TPS) but are microbial terpene synthase like (MTPSL), more closely related to fungal or bacterial terpene synthases. This type of terpene synthase only appears to exist in non-seed plants **(Jia et al., 2016)**, where they are involved in the production of monoterpenes and sesquiterpenes. Terpene products have been shown to have antimicrobial and antifungal

effects **(Asakawa & Ludwiczuk, 2018)**, thus their association with Marchantia's response to *C. nymphaeae* is quite logical, even though so far, they were only proved to be associated with herbivore deterrence **(Romani et al., 2020)**.

Most of the susceptible accessions (approximately 12) lack the genomic region under the whole peak, even in the 4 MTPSL genes (Figure 47). This shows the non-negligible effect of structural variation (here presence absence variation) on trait variation **(Della Coletta et al., 2021)**.

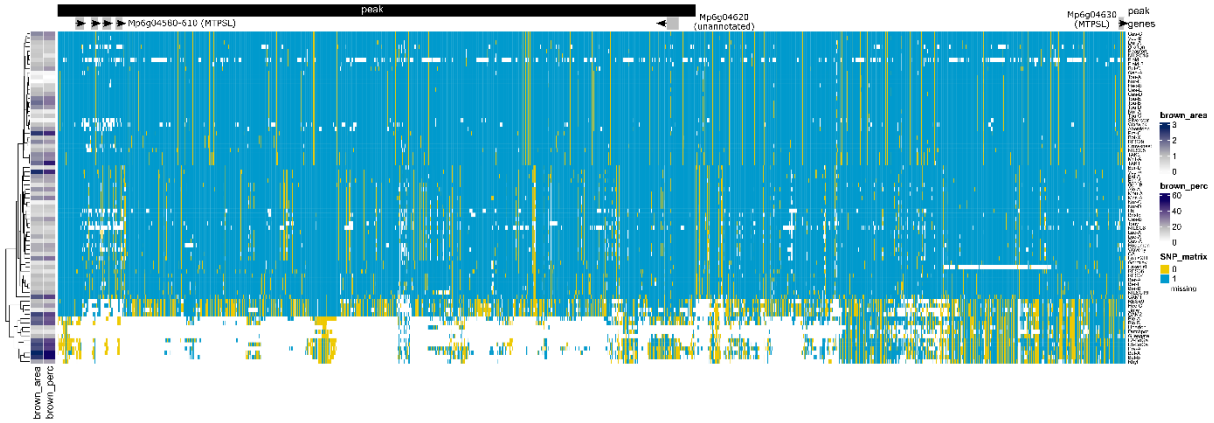


Figure 47: Detail of the alleles on the SNPs located close to the peak of the MTPSLs (chromosome 6). Most of the accessions with high susceptibility to C. nymphaeae map poorly in the region of the peak, suggesting a long stretch of deletion.

Some other peaks only pass the significance threshold for one of the studied phenotypes. A peak on chromosome 5 was only detected associated to the browning area and is surrounded by an Orotidine 5' decarboxylase (Mp5g15200), that usually participates in the synthesis of pyrimidine nucleotides, and an atypical kinase ABC1K (Mp5g15210), from a clade only found in algae, bryophytes and lycophytes, that has a common ancestor with the ABC1K12 mitochondrial family **(Lundquist et al., 2012)**, whose role is not well defined. Interestingly, this gene is under selective sweep in Marchantia (top 14% of genes with low values of Fay and Wu's *H*).

Another peak on chromosome 8 is due to the coexistence of two alleles on a long stretch of the genome, the minor allele being present in 34 accessions (*i.e* 44% of all accessions studied), 16 of them displaying severe browning symptoms. This genomic region is surrounded by an unannotated gene, and by a lipid droplet associated protein (Mp8g05270, MpLDPS2), that is involved in lipid droplet formation (Figure 48). Lipid droplets (not to be confused with bryophytes' oil bodies) are storage organelles, known for their role as nutrient reservoir in

seeds. They are also involved in abiotic stress response, which is their main role in non-seed plants. They do not only contain lipids but also enzymes allowing the production of specialised metabolites, like terpenes in bryophytes **(de Vries & Ischebeck, 2020)**.

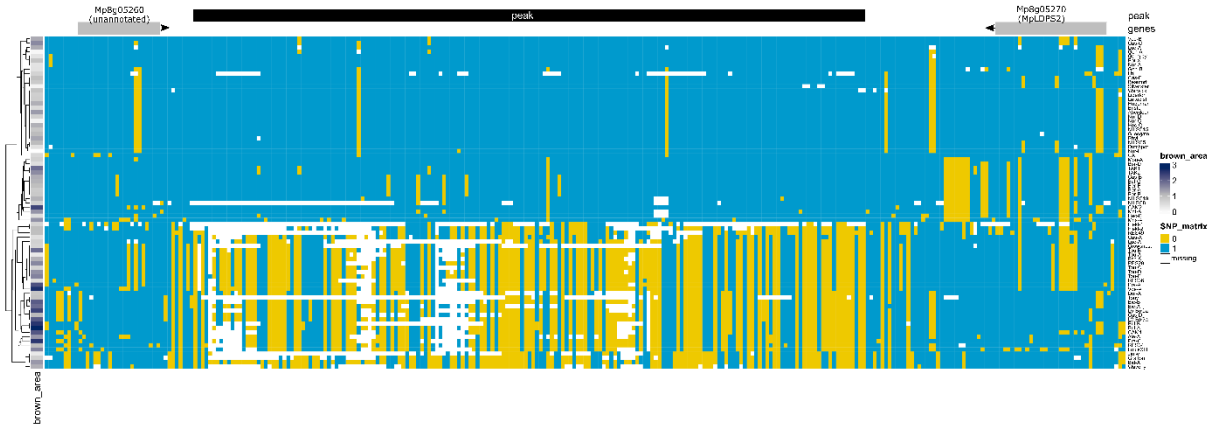


Figure 48: Detail of the alleles on the SNPs located close to the peak of the MpLDPS2 (chromosome 8). The minor allele stretch responsible for the signal is quite common in the population: 34 out of the 77 accessions phenotyped for the GWAS.

According to the PANTHER annotation, *M. polymorpha's* LDPS is bearing a 1,8-cineole synthase domain, that can transform geranyl pyrophosphate, the precursor of monoterpenes, in 1,8-cineole (also known as eucalyptol). After inspection of the phylogenetic tree of the MpLDPS2 gene on landplants (same strategy as the phylogeny for the GEA candidates, blast e-value of 10^{-20}), *M. polymorpha's* candidate appears to be orthologous to both the LDPS (AT3G19920) and the 1,8-cineole synthase (AT5G64230) gene of A. thaliana, that seems to have split during a seed plant specific duplication (Figure 49). This potentially adds another proof of the importance of terpene metabolites in *Marchantia*'s resistance to *C. nymphaeae*, this time coming from enzymes in the lipid droplet organelle.

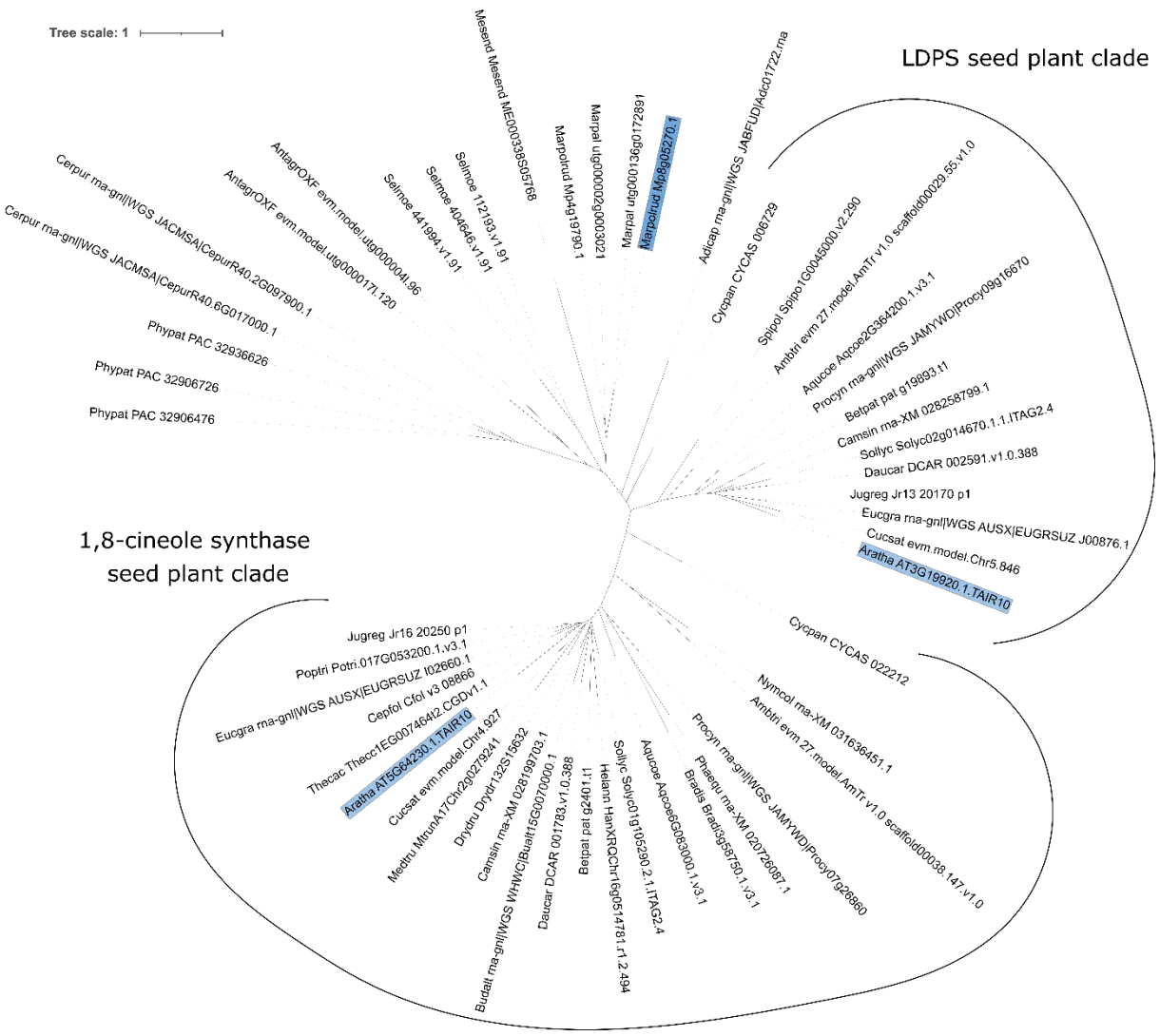


Figure 49: Landplant orthologs of the MpLDPS2 gene. Bryophytes lycophytes and ferns only have one gene whereas seed plants underwent a duplication leading to LDPS and 1,8-cineole genes.

The genome wide association on the percentage of thallus showing browning symptoms also lead to specific genomic regions. The first one is on the second chromosome, close to two unannotated genes and followed by a heavy metal associated protein (Mp2g25230), that is down regulated for various abiotic stresses (mainly related to heat and dark) and *P. palmivora* infection at 3dpi. One of these proteins has been shown to regulate the salicylate dependent immunity pathway **(Zschiesche et al., 2015)**, but the association signal seems to be closer to the unannotated genes.

The second peak is on the fifth chromosome and overlaps with a cluster of 3 peroxidases (Mp5g12-110, 120 and 130: POD93, POD94 and POD95) (Figure 50). Interestingly, according to peroxibase **(Passardi et al., 2007)**, these class III peroxidases belong to the same orthogroup as the one found in the GEA (POD128). This shows the importance of this family of peroxidase,

originating from a *M. polymorpha* specific family expansion, in its response to various stresses. The expansion allowed the genes to undergo different selective forces: POD95 is under strong conservative selection (very few SNPs) whereas POD94 and 93 are among the genes with the most balancing selection (top 2% of balancing genes). POD93 is down regulated in response to *P. palmivora* (4 dpi) and dark stress, whereas POD94 is up regulated in nitrogen deficiency (POD95 is not differentially expressed in any of the stress conditions covered).

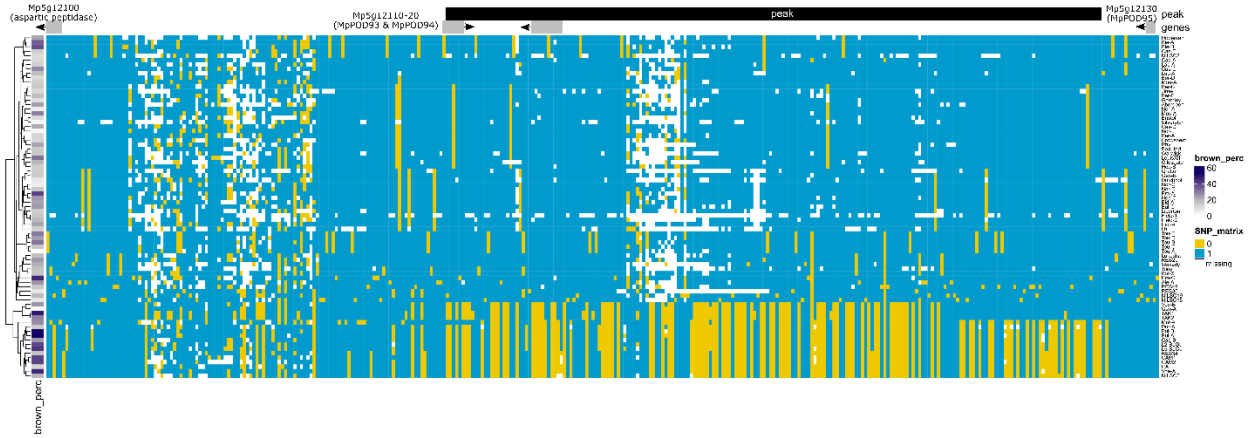


Figure 50:Detail of the alleles on the SNPs located close to the peak of the peroxidases (chromosome 5). The association signal is caused by a stretch of alternative allele present in 15 accessions that are among the most susceptible to the pathogen.

The GWAS on the browning phenotypes gave results that are a bit easier to interpret than the developmental GWAS. Most peaks are surrounded by well annotated and studied genes that can be linked to processes involved in plant defence. Interestingly, some candidates found here echoes the candidates found with the GEA, like the terpene synthases, the peroxidases or the ABC1K.

b) Focus on the terpene synthases

Since terpenes seems to have a key role in the response of *M. polymorpha* to pathogens (as well as in abiotic stresses with the MpIDS1), and since some of them (the microbial terpene synthases like) are non-seed plant genes, I tried to understand a bit more how this family is organised in *M. polymorpha*.

The terpene synthase landscape in M. polymorpha

All the proteins annotated as terpene synthases in the reference genome were blasted (blastp with an e-value of 10^{-10}) against the proteome (18 334 predicted proteins), to verify no candidate was omitted. This led to a list of 53 genes corresponding to various types of terpene

synthases. This gene list was blasted against itself (blastp with e-value filter of 1, to be able filter the similarity hits at later stages) and a similarity network was plotted using Cytoscape (Shannon et al., 2003), only representing the edges accounting for similarity e-values inferior to 10^{-25} (Figure 51).

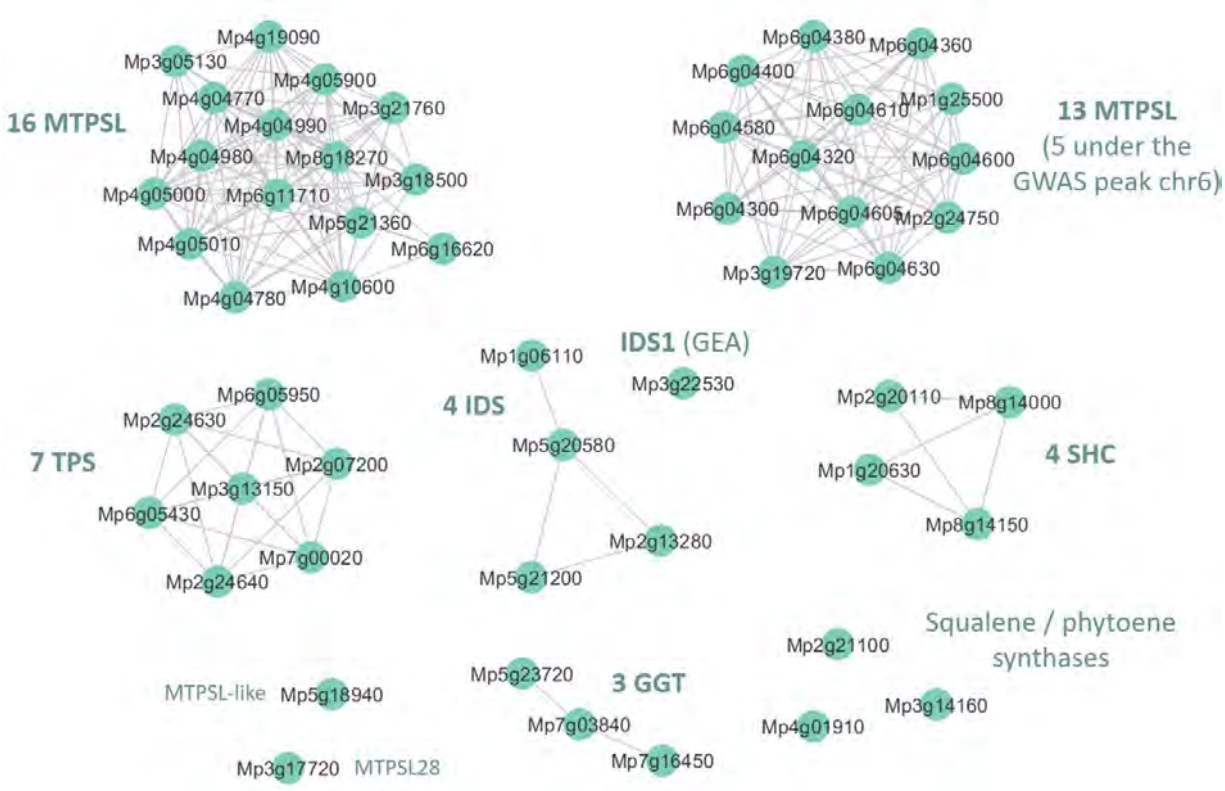


Figure 51: Sequence similarity network of M. polymorpha's terpene synthase genes. The bryophyte counts a majority of microbial terpene synthase-like (MTPSL) with a family of 16 genes and a family counting 13 genes. TPS stands for plant terpene synthase, IDS for isoprenyl diphosphate synthase, SHC for squalene hopene cyclase, GGT for geranylgeranyl transferase.

This sequence similarity network clearly shows that most of *Marchantia*'s terpene synthases are of microbial origin (29 MTPSL divided in two families). On the other hand, 7 terpene synthases are plant terpene synthases that usually synthesise diterpenes in non-seed plants (three of Marchantia TPS have proven diterpene activity (Kumar et al., 2016)) and were derived to also synthesise mono and sesquiterpenes in seed plants (Jia et al., 2016). Five others are isoprenyl diphosphate synthase, enzymes responsible for the production of terpene precursors. The GEA candidate IDS1 is a farnesyl diphosphate synthase (FDS) that synthesises the precursor of sesquiterpene, whereas the four other IDS are annotated as geranylgeranyl pyrophosphate (GDS), allowing the production of the monoterpene precursor. Four other proteins are annotated as squalene-hopene cyclases (SHC), that converts squalene, a linear

triterpene, into cyclic triterpenes like hopene or lanosterol (the precursor of sterols and steroids) and three as geranylgeranyl transferases (GGT) which are post translational modifiers of proteins. Finally, there are three dissimilar squalene/phytoene synthases that catalyse the condensation of farnesyl diphosphate (squalene synthase, SQS) or geranylgeranyl diphosphate (phytoene synthase, PSY) and two MTPSL named following research work **(Kumar et al., 2016)**, but with no functional annotation. The role of these enzymes in the terpene synthesis pathway is detailed in Figure 52.

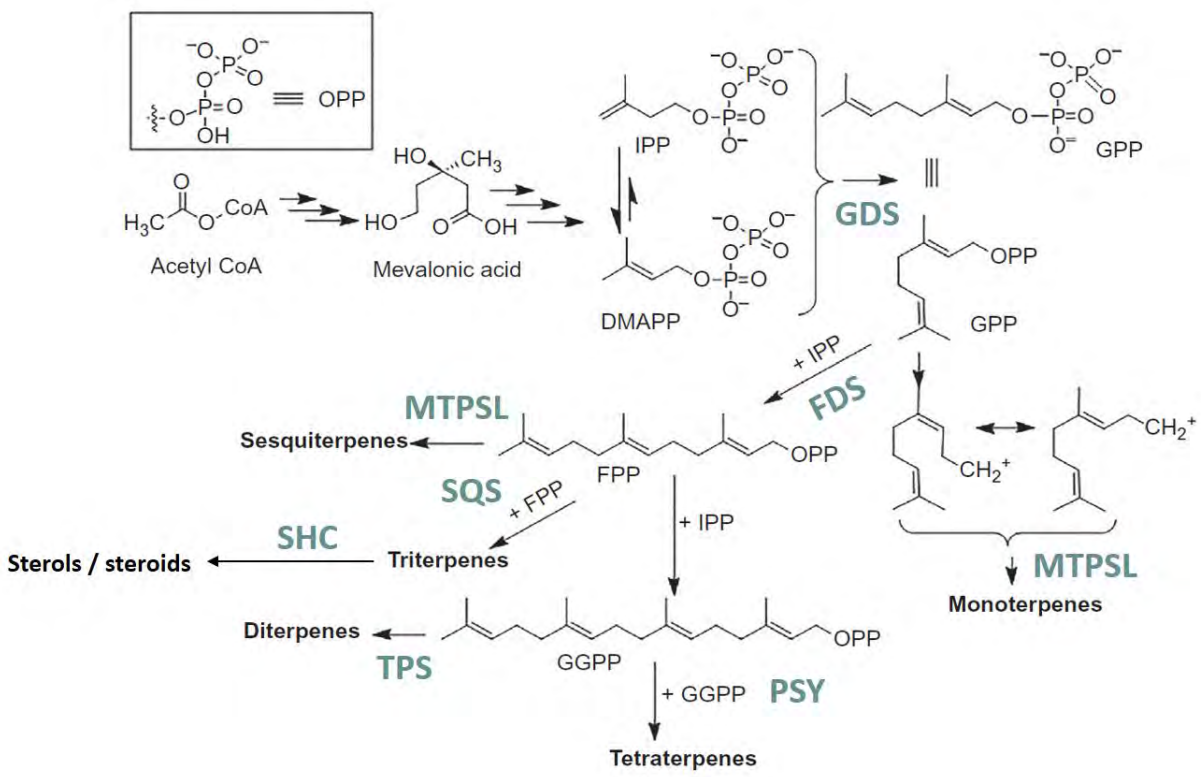


*Figure 52: Role of the terpene synthesis related enzymes found in Marchantia in the terpene biosynthetic pathway (adapted from **(Habtemariam, 2019)**. MTPSL: microbial terpene synthase like, GDS: geranyl diphosphate synthase, FDS: farnesyl diphosphate synthase, SQS: squalene synthase, PSY: phytoene synthase, SHC: squalene hopene cyclase.*

The origin of Marchantia's MTPSL

Terpenes synthases exist in fungi, bacteria and plants, but despite their similarity in catalytic function, reaction mechanism and structure, typical plant terpene synthases share very little sequence similarity with microbial terpenes synthases, suggesting different evolutionary origins. Given this clear difference between plant and microbial terpene synthases, microbial terpene synthases-like, found in non-seed land plants (liverworts, mosses, hornworts, lycophytes and monilophytes), probably originate from horizontal gene transfers **(Jia et al., 2016)**. *Marchantia* bears two different families of MTPSL, which leads to wonder if these MTPSL

originated from two distinct gene transfer from microbes. In order to answer this question, phylogenetic analyses were conducted for the MTPSL from each family (that I will call 16 MTPSL and GWAS MTPSL). The MTPSL genes were blasted against a database with viridiplantae genes (SupData2.4 sheet 1), a database with non-angiosperm transcriptomes from the 1KP initiative (One Thousand Plant Transcriptomes Initiative, 2019), a database with fungal genomes from MycoCosm (Grigoriev et al., 2014), a database with algae genomes (SupData2.4 sheet 2) and the non-redundant (nr) database from the NCBI (blastp with evalue 10^{-5}, 2000 target maximum). The resulting proteins were aligned with muscle5 (default parameters), trimmed with trimal (`-gt 0.4`) and the phylogenetic analysis was conducted with IQtree (`-alrt 10000 –B 10000 –T AUTO` options).

The 16 MTPSL family appear to have orthologs in fungi and bacteria, it is therefore complicated to conclude on the initial donor, from which the MTPSL gene was passed to the plant (Figure 53). The plant clade that bears orthologs of the Marchantia genes are liverworts, mosses but also lycophytes. The presence of lycophytes, a clade of vascular plants, in this tree would suggest that the horizontal gene transfer (HGT) of this type of MTPSL occurred in the common ancestor of land plants, and was then lost in several clades (hornworts, and in the ancestor of ferns and seed plants). Most of the lycophyte proteins come from transcriptomes (mostly from lycopodiales species like *Huperzia lucidula*, *Huperzia selago*, *Lycopodium deuterodensum*, or *Pseudolycopodellia caroliniana*), but also 7 gene models from the genome of *Lycopodium clavatum*. The genomes of *Selaginella lepidophylla* and *Isoetes taiwanensis* did not display any ortholog of the MTPSL in their proteomes. Therefore, to verify the extent of the presence of this gene in lycophytes, their nucleotide sequences were screened with a tblastn (e-value 10^{-3}). This resulted in the Fgenesh annotation of one gene model in *Isoetes* and 3 in *Selaginella*. This strengthens the hypothesis of the ancestral origin of this MTPSL in landplants.

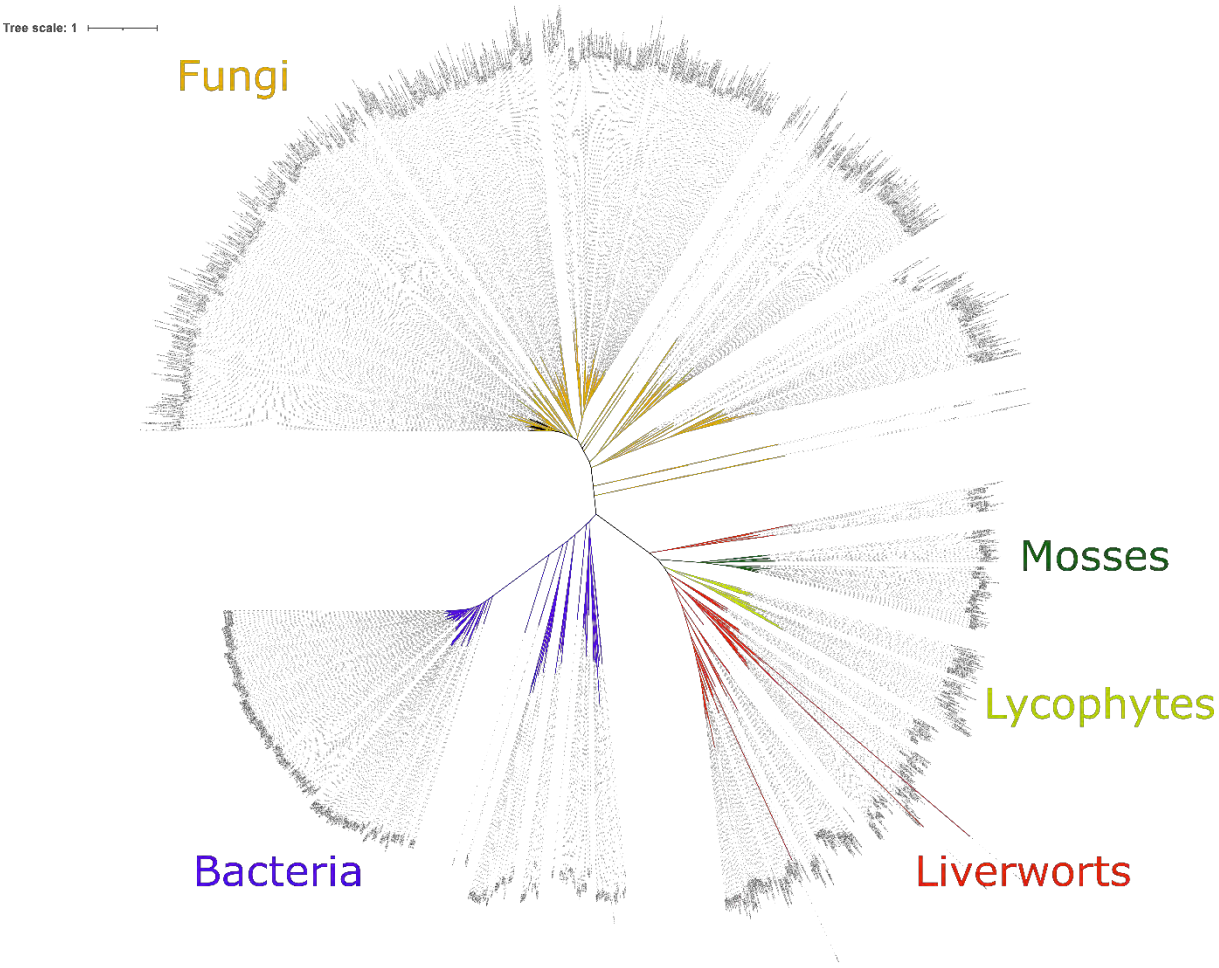


Figure 53: phylogenetic tree of the 16 MTPSL family. This family originated from a horizontal gene transfer from either a fungi or a bacterium, and is present in mosses lycophytes and liverworts, suggesting a gain in the common ancestor of land plants.

The GWAS MTPSL family was found in fungi (Ascomycota and Basidiomycota), and in liverworts (10 out of the 30 included in the databases, among which six genomes, most of which are Marchantiales) and hornworts (7 out of the 10 included in the databases, among which three genomes). This suggests an HGT from the common ancestor of the Dykaria fungi to the common ancestor of the bryophytes, followed by the loss of the gene in mosses.

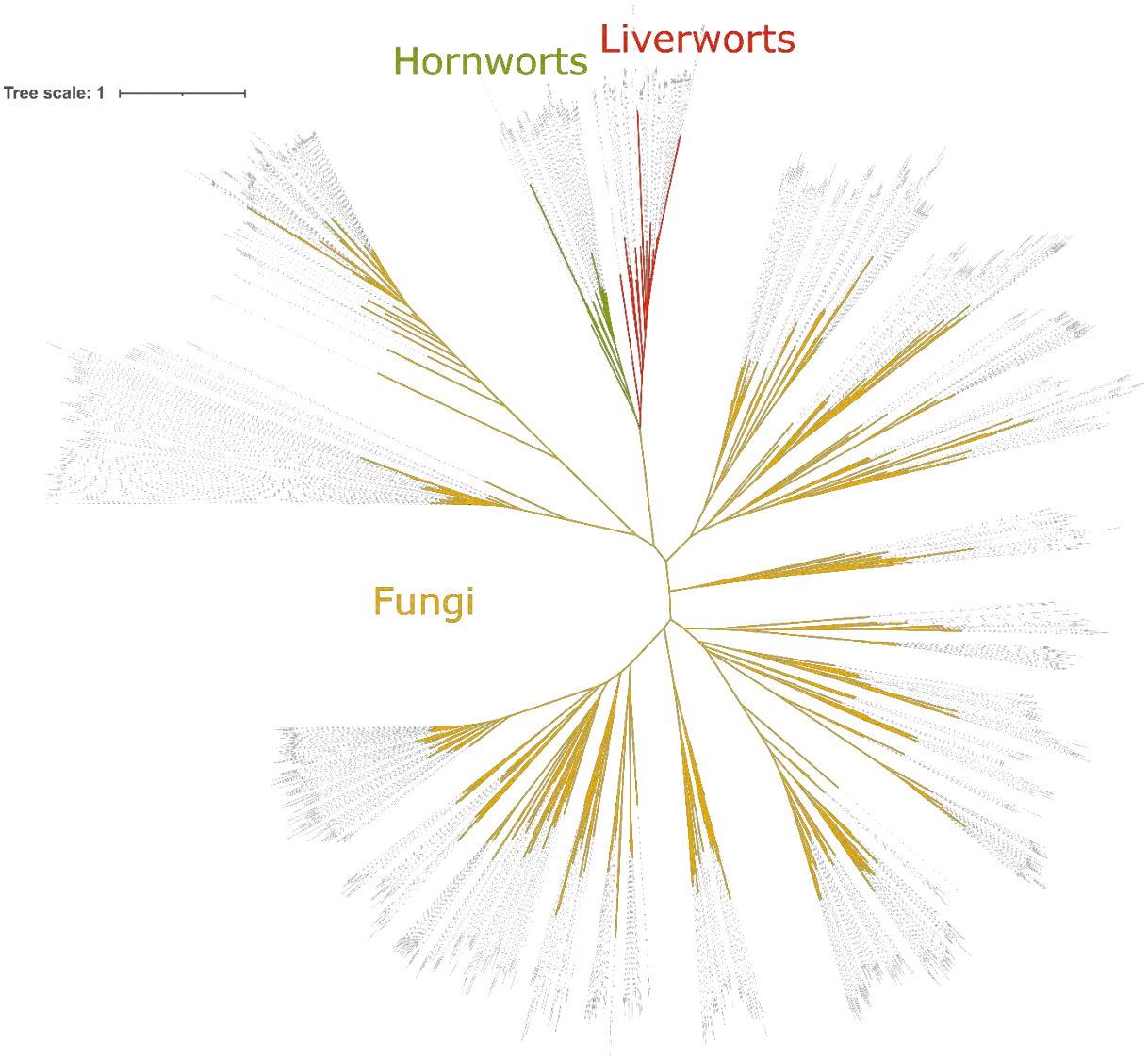


Figure 54: phylogenetic tree of the GWAS MTPSL family. This family originated from a horizontal gene transfer from a fungus, and is present in hornworts and liverworts, suggesting a gain in the common ancestor of bryophytes.

Since the HGT is more recent than for the other MTPSL family, it is possible to observe how the MTPSL have spread in the bryophyte genome: 10 out of the 13 genes are located on the same chromosome, in a 377 Kb stretch, under two gene clusters. The signature of the same transposon is found flanking the genes, suggesting this was their means of duplication (Figure 55). This may be the cause of the structural variability among the accessions in this region, observed on the SNP matrix (Figure 47) and previously noticed on an accession not present in our collection (Takizawa 2021).

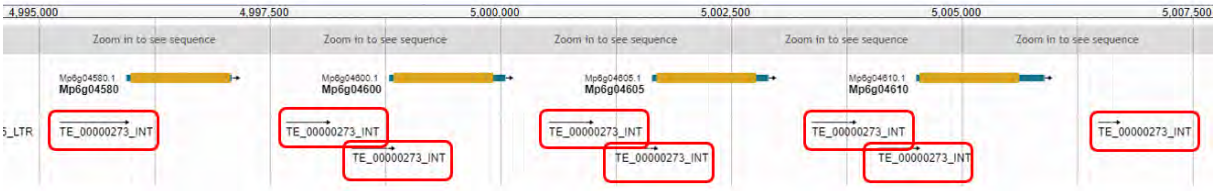


Figure 55: extract of M. polymorpha's genome browser, showing the transposon signature flanking all the genes under the GWAS peak.

The other MTPSL family (the 16 MTPSL) have probably been duplicated after their transfer in *Marchantia*'s genome, even though the transposon signature is not as clear. The duplication on these genes probably relaxed the selective pressure, and three MTPSL from this family are among the genes with missense mutations that are almost fixed in the population (cf chapter 1): (Mp4g04770, Mp4g04980, Mp4g05000).

These two MTPSL families come from two distinct horizontal gene transfer, at different times of the land plant evolution. One of these events may find its origins in the common ancestor of land plants, which is consistent with the HGT burst observed during plants terrestrialisation **(Ma et al., 2022)**, whereas the other seems to be a bryophyte specific gene, consistent with the weak link model that hypothesises that organisms with exposed reproductive structure or early developmental stages tend to undergo more HGT **(Huang, 2013)**. Given their orthologs, the GWAS MTPSL family corresponds to the group I (suggesting that the donor was actually bacterial) and 16 MTPSL family to the group III of MTPSL, described by Jia and collaborators **(Jia et al., 2016)** based on 1KP transcriptomes. Our finding also corroborates the groups delineated by Kumar et al. **(Kumar et al., 2016)** based on sequence similarity network between some MTPSL from Marchantia and a variety of terpene synthases. Our GWAS MTPSLs corresponds to their 2-methyl isoborneol synthase group III (monoterpene synthase) and our 16 MTPSL family corresponds to their pentalenene synthase group II (sesquiterpenes synthase). Compared to this study, we identified new genes in both families, allowing a complete description of the MTPSL landscape in *M. polymorpha*.

III) Concluding remarks on the genome wide association studies (GWAS on the response to *C. nymphaeae* and GEA)

The genome-wide association study of the response of *M. polymorpha* to the fungal pathogen *C. nymphaeae* highlighted some interesting candidate genes. First, data from the non-inoculated condition was used to investigate the genetic bases of developmental variation in *M. polymorpha*. For some significantly associated regions, it was difficult to determine which gene was causal and how it could be involved in development. Nevertheless, some function stood out: cell wall plasticity (Expansins, UDP glucosyl transferases), lipid metabolism (GDSL lipases), and some more general ones like transcription factor and F-box protein.

For the GWAS of the response to the pathogen, different candidates were found. Among the most promising ones are a thioredoxin, a receptor-like kinase likely probably involved in brassinosteroid signalling (BSK), a cluster of terpene synthases, an ABC1K protein, a lipid droplet associated protein and class III peroxidases. Some functions are similar to the ones found deregulated by the infection of *M. polymorpha* by *Phytophtora palmivora*: the up-regulated peroxidases (with the PR9 family in the *P. palmivora* study), transcription factors, terpene synthases (down regulated in response to *P. palmivora*), receptor like kinases (up-regulated) **(Carella et al., 2019)**, and might be key functions of the core immune response of *M. polymorpha*.

Different commonalities were found between the GWAS on response to pathogens and the GEA: they share class III peroxidase candidates from the same orthogroup, ABC1K protein and terpenes synthase genes. This outlines some common basis of resistance against both biotic and abiotic stresses in *M. polymorpha*.

The wealth of evidence leading to the terpenoids pathway made it of specific interest. Two types of genes from this pathway seem involved in the response of *Marchantia* to *C. nymphaeae*: MpLDPS2 that enables to produce a monoterpene, and a cluster of microbial terpene synthase-like, that are also involved in the last steps of the mono or sesqui-terpenes biosynthesis. Another gene from this pathway, MpIDS1 a farnesyl diphosphate synthase, was also detected in the GEA. This highlights the importance of terpenoid compounds and notably sesquiterpenes in the response of *M. polymorpha* to various stresses.

The similar candidate genes shared by the GWAS and the GEA analyses suggest a potential crosstalk between biotic and abiotic stress response pathway. This is exemplified by the NLR found with the GEA (NBS-LRR11), that is one of the few *M. polymorpha's* CNL. NLR are usually involved in the detection of pathogen effectors and response to infection, but this one was found by an association analysis on environmental variables, and has been found up-regulated under abiotic stress conditions as well as biotic stresses: infection by *P. palmivora* a 2,3 and 4 dpi, as well as nitrogen deficiency and two other nitrogen deficiency stresses combined with cold or salt (Carella et al., 2019; Tan et al., 2023). We could hypothesize that the NLR is up regulated by nitrogen deficiency because it could help the plant to shape its microbiota and by this mean get a better access to nitrogen. Indeed, even though *M. polymorpha* does not associated with arbuscular mycorrhizal fungi, it can associate with classical plant-associated beneficial bacteria like *Methylobacterium* species that helps the plant with nitrogen fixation (Alcaraz et al., 2018).

This interplay between immunity and other stresses in plant has been well documented on many immunity actors, like pattern recognition receptors, secondary metabolites and hormones (Saijo & Loo, 2020). The interactions between biotic and abiotic stresses can have diverse outcomes. Some NLRs can increase the plant sensitivity to abiotic stresses, because of a trade-off between stress responses (L. Yang et al., 2021). For instance, the NLR ACQOS disables osmotolerance when it is functional (Ariga et al., 2017). Meanwhile, other NLRs offer combined resistance and sensitivity, like the broad spectrum disease resistance CNL ADR1, that confers drought tolerance, but also thermal and salinity stress sensitivity, through the negative regulation of ABA signalling (Chini et al., 2004).

It should be noted that some peaks of the genome-wide association studies did not overlap with predicted gene models, but they were frequently located in the putative regulatory regions of some genes, suggesting a potential role for cis-regulation in the natural variation of responses to stress. Different types of transcription factor were also found as candidate genes in both genome wide association studies (trihelix TF, lateral organ boundary TF, Myb TF, bHLH TF). As mentioned in the introduction, the importance of regulatory networks in the natural variation of stress responses or changes in environmental conditions has already been observed in GWAS in other plant species, such as *A. thaliana* (Frachon et al., 2018). TF

importance has also been proven in global studies of plants (and particularly crops) response to specific stresses, such as drought (Joshi et al., 2016).

Chapter III

The first pangenome in a bryophyte: construction and exploration

I) Building *M. polymorpha's* pangenome: a trial-and-error process

Candidate genes and genomic regions linked to *Marchantia polymorpha* ssp. *ruderalis* adaptation to biotic and abiotic environment were identified via genome-wide association study, gene-environment association and, indirectly, by the study of selection signatures on genes. However, single nucleotide polymorphisms are not the only variants behind genetic adaptation. Structural variants also have a non-negligible effect in adaptation processes, since they are a source of diversity between accessions. They have been shown to explain some of the missing heritability highlighted in GWAS analysis **(Y. Zhou et al., 2022)**, which is consistent with the missing data surrounding the GWAS and GEA candidates, that may be explained by a poor mapping of the accessions with structural variation in the region (Figures 29, 33, 47). To be able to get a broader view of *M. polymorpha* genetic diversity and to spot potentially adaptive genes not necessarily present in the reference genome, we initiated the construction of the first pangenome in a bryophyte species.

1) First trial: the map to pang strategy

The initial approach we adopted for the pangenome construction was the map-to-pan, a classical strategy for the treatment of short reads sequences **(Bayer et al., 2020)**. It consists in the iterative assembly of the non-mapping sequences of the different accessions. This pangenome construction was carried out by Hélène San Clemente, from the bioinformatics service of the LRSV, following the operations detailed in the Figure 56.

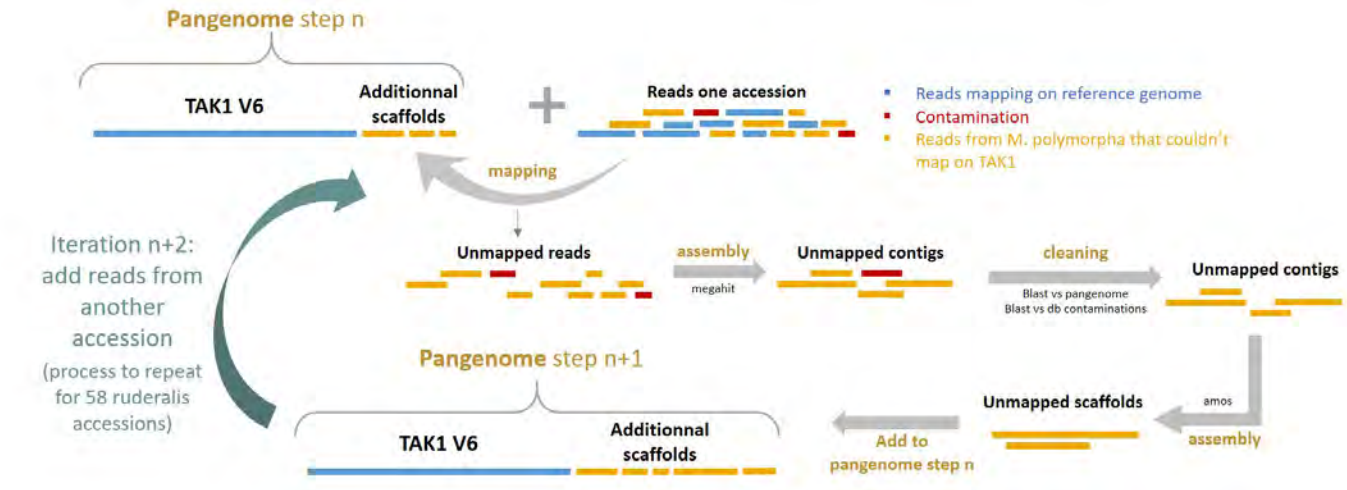


Figure 56: Details of the pipeline used for the iterative assembly of M. polymorpha ssp. ruderalis pangenome for 58 accessions. For each accession, the same loop of operation repeats itself: the reads not mapping on the

reference genome (and on the additional scaffolds already build with previous accessions) are retrieved, cleaned and assembled. This leads to a pangenome composed of the reference genome and additional scaffolds not present in the reference but present in some of the accessions that were processed.

For each iteration on an accession, the operations are the following ones:

- the reads from the accession are mapped on the pangenome generated at the previous iteration (composed of the reference genome from the TAK1 accession, and of some additional contigs).

- the non-mapping reads are assembled and then blasted (blastn) against the pangenome and against the nucleotide database from the ncbi (nt). If there is a hit on the pangenome, or for something from the nt that is not a Marchantia sequence, the contigs are discarded. The contigs that do not hit against anything in the nt are kept.

- the remaining contigs are deduplicated (if two or more contigs correspond to the same genomic region, only one will be kept) and assembled another time. The initial reads are mapped on these contigs to control for their sequencing depth, the contigs with a sequencing depth lower than half of the mean sequencing depth of the pre-existing pangenome being discarded.

- the contigs are added to the pre-existing pangenome, leading to a new version on which all these steps can be repeated.

This procedure led to a pangenome with 94 239 additional scaffolds representing 244 Mbp. This is quite surprising considering that the reference genome is about 230 Mbp. Considering the size of the pangenome obtained in other plants (Bayer et al., 2020), it seemed quite unlikely that adding less than 60 accessions multiplies by two the genomic content of *Marchantia polymorpha ssp ruderalis*.

To explore this pangenome a bit more and try to understand the reason for this substantial size change compared to the reference, the protein coding genes were predicted. The additional scaffolds were annotated with BRAKER2 pipeline (Brůna et al., 2021), with the `--prot_seq [TAK1_reference_proteome] --prg gth --trainFromGth --species pangenome_marpolrud_w_TAK1 --softmasking --gff3 --cores 8` options, meaning that the protein prediction is based on the proteins already predicted in the reference genome, considered as a “closely related species”. Hints of prediction were generated by

GenomeThreader (`--prg gth` option) and used to train AUGUSTUS (`--trainFromGth` option) to structurally annotate the genome. This led to the prediction of 115 642 genes, that seemed again very high compared to the 18 334 genes predicted in the reference genome. Those predicted genes were then examined for their presence absence among accessions, by using the SGSGeneLoss software **(Golicz et al., 2015)** that calculates the read coverage on gene exons for each accession, to know if the gene seemed present or not in a given accession. I screened the pangenome for the 76 *ssp ruderalis* accessions for which we had sequencing data at that time, even though only 58 of them had sequences of good enough quality to be used in the pangenome construction. From these data I could determine which genes were shared by all accessions and which genes were only present in a few accessions (Figure 57).

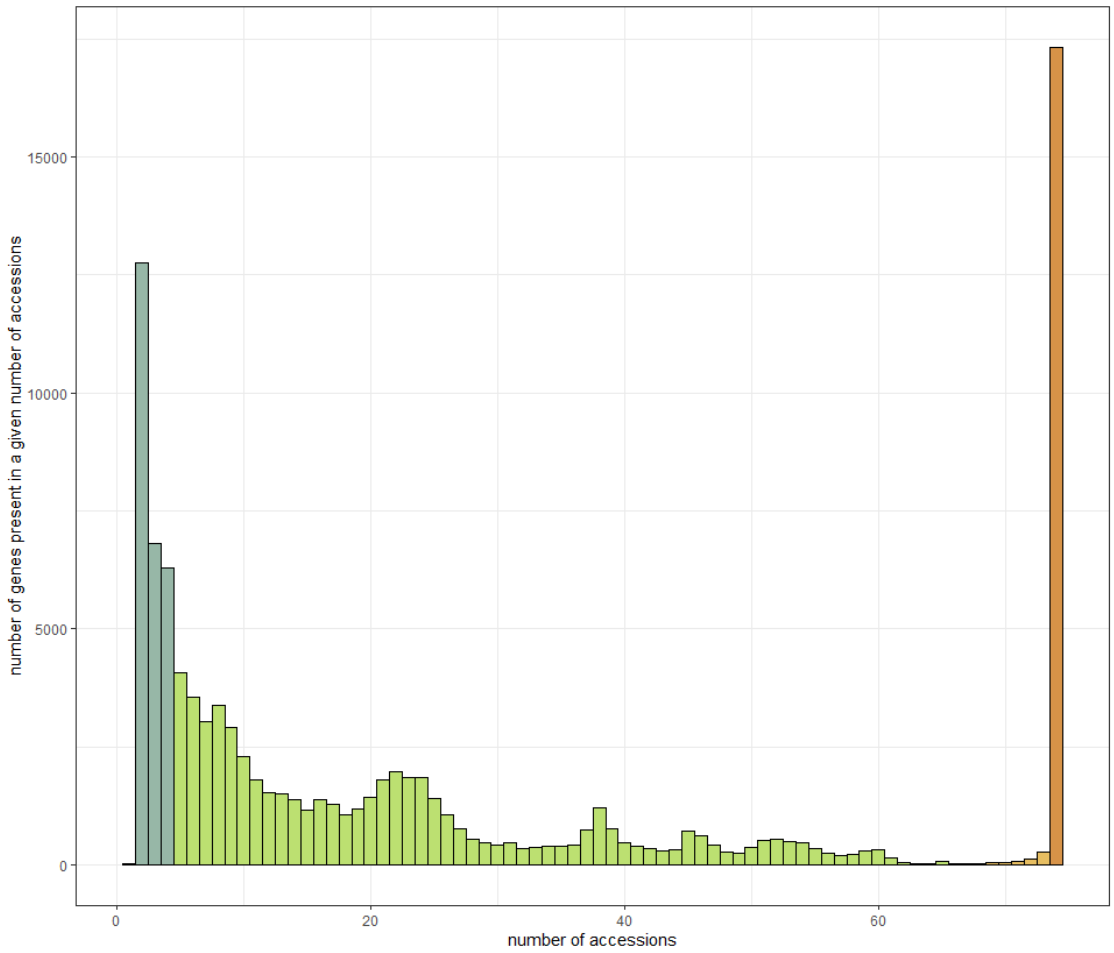


Figure 57: Distribution of the genes predicted in the pangenome (reference genome initial annotation and additionnal scaffolds annotation with BRAKER2), depending on their presence in a given number of accessions. The orange bar represents genes shared by all accessions, and green genes represent genes only present in some accessions, with the grey green bars on the left being genes shared by a very limited number of accessions.

The distribution displays on the right side the core genes, shared by all accessions, and on the left side the very rare genes, shared by few accessions. In this distribution there is a surprisingly important number of rare genes, even more than core genes. This is probably due to shared contaminations in sequences from a few accessions, that were not discarded by the repeated rounds of cleaning. The presence of contaminated contigs is visible on the Figure 58, considering the distribution of the contigs depending on their GC content and their length. The average GC content for *M. polymorpha* is around 42% (according to the reference genome data), and there is clearly a bimodal distribution of the contigs GC contents: some of them are around 42% GC, whereas a non negligeable number of them are around 55-65% GC. Moreover, a lot of long contigs (corresponding to high values on the y axis) have GC contents deviating from the expected value for *M. polymorpha*, which suggests that long stretches of contaminating sequences seem to have been assembled.

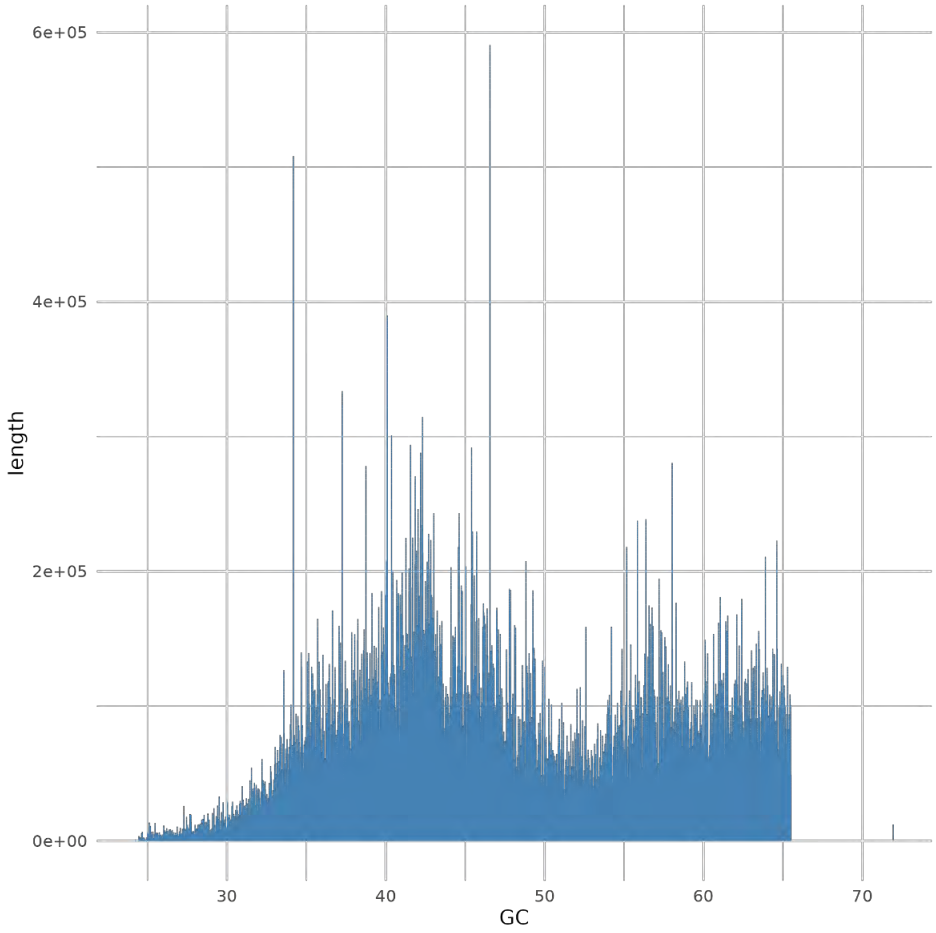


Figure 58:Distribution of the additional contigs of the pangenome, depending on their GC content, and with the information of their length in bp. The Gaussian around 42% GC could represent M. polymorpha contigs, but the ones deviating too much from this expected GC content are probably contaminations.

To obtain another proof of the presence of contamination, I compared the genes identified as present in the CA accession (based on the mapping of short reads) to the genes present in the long reads assembly of the same accession, by a tblastn of the predicted genes on the long reads genome. Only 2 505 of the 16 317 pangenomic genes predicted to be in CA were actually in the long reads assembly. A GO term enrichment on the remaining genes indicated terms linked to bacterial flagella, viral capsids, or secretion system, confirming the significant contamination of the additional contigs.

I tried to clean the contigs using a blast of their genes against the non-redundant protein database of the ncbi (nr), to see if I could retrieve the contigs bearing multiple eukaryotic genes. But only 1.17% of the genes were in this case and they were scattered in different contigs. It was therefore hard to develop a strategy to rescue the "clean" sequences of this pangenome.

After discussing about the process of the sequencing, it appeared that most of the accessions had been sequenced after cultivation in soil, and not in axenic culture. The DNA extracted from these plants must contain a substantial amount of DNA from all the microorganisms that were in their close vicinity. Conscious of this issue, we tried another approach to clean the data and to be able to build a pangenome.

2) Second trial: preliminary k-mer cleaning

The cleaning processes used in the pangenome construction pipeline did not consider enough the contamination in our data. I therefore tried to clean the reads library themselves, by filtering them based on their k-mer and GC content. Cleaning before the assembly of the additional genome allows to have less data to process, but also to have a better guarantee that no chimeric assembly will be formed between different types of contaminant sequences or plant DNA and contamination.

The principle of k-mer distribution study is basically to decompose the reads in "words" of a given length (the k-mers, of length k), and to count the occurrence of each given word in the sequence. This leads to a gaussian distribution and allows to spot k-mers that deviate from this distribution (that are either too rare or appear too many times) and may therefore come from a contamination. To implement this, the KAT software **(Mapleson et al., 2017)** was used. The reads library of each accession were processed with Trimgalore! v 0.6.5 and then as an input in KAT with the gcp option and a k-mer length of 27 (default parameter), in order to produce plots

of the distribution of the 27-mers depending on their frequency and on their GC content. This allowed to detect at the same time GC bias and k-mer distribution bias. These plots were then used to manually determine GC and k-mer frequency thresholds that would allow to clean the reads. The k-mer within these thresholds, that were likely to originate from Marchantia's DNA, were used to clean the reads: only the reads with 75% of their sequence covered by this type of k-mers were kept. An example of this cleaning process is displayed in the Figure 59.

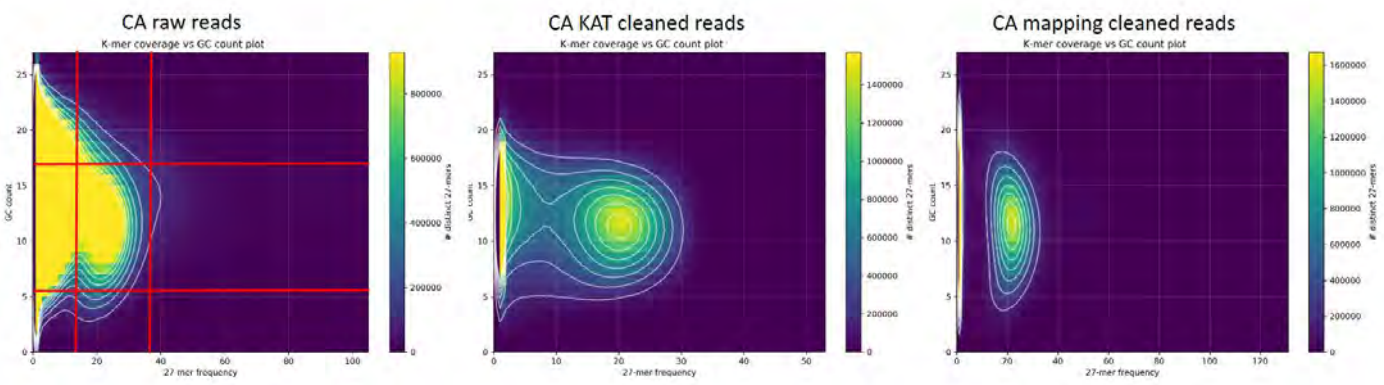


Figure 59: Distribution of the 27-mers from the CA accession, according to their frequency of occurrence in the libraries and their GC content. On the left is the distribution on the raw reads from CA, with the red lines showing the thresholds taken to clean those raw reads. In the middle is the distribution of 27-mers on reads after cleaning depending on the k-mer content of the reads. On the right is the distribution of the short reads from CA cleaned by mapping on the long reads assembly, that shows the true distribution of non-contaminated reads.

The Figure 59 clearly shows that the cleaning based on k-mers allows to discard a great part of the contaminations, but that there is still a non-negligible amount of abnormal k-mers compared to the clean set of reads from CA (gotten by mapping on the long reads assembly, as a test because it is the only accession for which such cleaning can be done).

This method was applied on the illumina sequencing from 69 accessions. For 19 of them, there was not enough reads after the cleaning for them to be used in the pangenome construction. The new construction was therefore carried out on 57 accessions, with a similar method to the first trial, but conducted in parallel instead of iteratively. The cleaned libraires were therefore assembled separately for each accession with MEGAHIT **(D. Li et al., 2015)**, and the resulting contigs processed with the genome assembly evaluation tool QUAST **(Gurevich et al., 2013)**. This tool performs multiple analysis, among which the mapping of the assembly on a reference genome. This allowed to spot the non-mapping contigs, selected as contigs of minimum 1000 bp, with at least 70% of their total length not mapping to the reference genome and maximum

3 distinct unaligned fragments. The unaligned contigs from all accessions were pooled and deduplicated with CD-hit **(Fu et al., 2012; W. Li & Godzik, 2006)**, leading to a total of 36 534 contigs potentially belonging to *M. polymorpha's* pangenome. These contigs were added to the reference genome and annotated with BRAKER2 as it was done for the previous pangenome. To do a quality control, the 68 553 genes were blasted against the NCBI nr database with DIAMOND **(Buchfink et al., 2021)**, with maximum 10 hits and an e-value of 10^{-1}. The proportions of hits of all the genes from the additional contigs was again clearly biased towards bacterial contamination (Table 4)

Table 4: Details of the hits for all the genes present in the additional contigs, in the different domains of life, also including viruses. The majority of correspondences with genes in the nr database are again with bacterial genes, showing that this version of the pangenome is still very contaminated.

Category	Number of hits	Percentage of total hits
Archeae	531	0,07%
Bacteria	753708	98,74%
Eukaryote	6863	0,90%
Other	27	0,003%
Unclassified	33	0,004%
Viruses	2187	0,29%

By looking more specifically into the blast hits, it appeared that 178 genes are similar to Viridiplantae genes. Those 178 genes are scattered in 152 contigs that contain a total of 297 genes. Therefore, there was once again very few additional genes that did not seem to originate from contamination, and these genes were scattered on fragmented contigs, among other genes that could originate from contamination. Given the low number of reliable genes found, and their patchy repartition, it seemed complicated to trust this new version of the pangenome as an accurate representation of the genetic variation in *Marchantia polymorpha*. After these two unsuccessful trials, we reconsidered the type of information we wanted to get with this pangenome and went for a construction strategy that would allow to focus on the cleaning of genes more easily.

3) Final idea: a gene-oriented pangenome

Since the idea of a linear pangenome of *M. polymorpha* seemed compromised by the nature of the data, we then tried a gene-based approach, that could simplify the cleaning processes by reducing the task: we did not look into the whole sequences anymore but only assessed

whether the coding sequences could come from contamination or not. This approach did not allow to detect all structural variants (particularly non-coding sequences) and to anchor them to the genome, but this type of analysis would have been complicated anyways considering that most of our data consisted of short reads sequencing. Focusing on the genes mainly allowed to detect the presence/absence variation of coding sequences across the accessions, and to find new genes that may not be present in the reference genome.

De novo assembly and structural annotation of accessions' sequences

This pangenome construction strategy was based on *de novo* assembly of all the genomes, contrary to the two previous strategies that were based on the mapping of sequences on the reference genome. The assembly of the short reads genomes (TAK1, CA and BoGa) had already been performed, I therefore assembled all the genomes sequenced by Illumina. The short reads of the 135 accessions were first processed with Trimgalore! v 0.6.5 **(Krueger et al., 2021)** to remove adaptors and control quality (-q 30 --length 20) and then assembled with megahit v1.1.3 **(D. Li et al., 2015)** with default parameters. The assemblies were then cleaned by discarding contigs that had abnormal features. The filters used to discard the contigs were the following:

-The contig length should be superior to 500 bp, the odds of predicting a gene in a smaller contig being very small

-The general GC content should be inferior to 55%. This value was chosen taking into account the the long reads assemblies global GC content of around 42% and the GC content distribution of short reads assemblies, to try to be just stringent enough without discarding to many potential *M. polymorpha* contigs. This allowed to discard some contigs that may belong to other organisms, but for some contaminated sequences this was not stringent enough to discard most of the contamination (Figure 60)

-The contig should have hits when blasted against the nucleotide (nt) database from the NCBI but no bacterial one (blastn with blast-2.13.0+, e-value 10^{-3}, maximum 10 target sequences). This filter could result in the loss of recent horizontal transfer or specific genes from *M. polymorpha*, but the contamination of some data was so important that the objective was to reduce drastically the amount of contaminated sequences before any other step of the pangenome construction.

The details of the outcomes of the filtering steps in terms of contig number and total base pair number for each accession is available in SupData3.1. The filtering resulted in more or less drastic reduction in sequence, depending on the accessions. Some of the cleanest like Withies lost 40% of their contigs, representing only 5% of the total base pair sequenced, whereas other accessions like NILSC21 lost 99% of their contigs, representing 90% of the total base pair sequenced. For some accessions the conservation of a great proportion of the contigs was actually a sign of the failure of the cleaning: RES40 lost only 20% of its sequenced data, but the protein prediction lead to a low BUSCO completeness score of 30% (Figure 60).

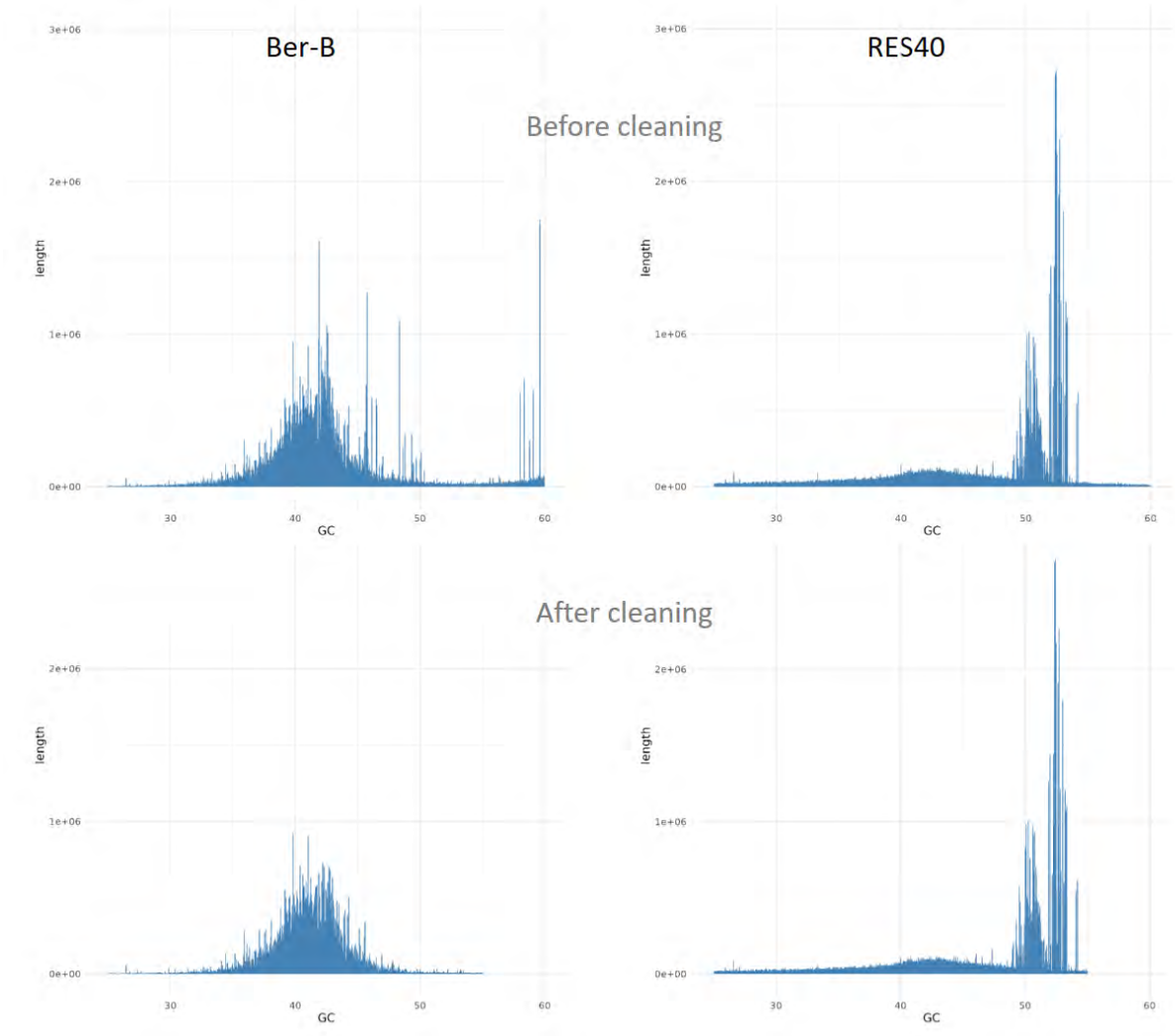


Figure 60: GC and length profile of contigs for two accessions, before and after the cleaning processes. For Ber-B (on the left), the cleaning process allowed to discard most of abnormal contigs. The cleaning was not so successful for RES40 (on the right), for which the GC filter was too laxist to discard the long contigs with high GC content. The blast cleaning did not seem to change the contig profile either.

The outcome of the assembly and cleaning of the short reads was heterogenous for the different accessions, which is logical given the heterogeneity of extraction and sequencing

processes. Post cleaning, the N50 of the assemblies, were comprised between 8 135 and 78 628.

After this cleaning, the remaining contigs of each accession were structurally annotated. The assemblies were softmasked with Red (Girgis, 2015), and the structural annotations were conducted using the BRAKER2 pipeline (Brůna et al., 2021), with the `--epmode --softmasking --gff3 --cores 1` options. Only one core is used to homogeneously process all the different parts of the fragmented assembly on the same core. The epmode is used to predict coding sequenced with the help of proteins from distant organisms (at least a bit more distant than with the Gth mode). The ProtHint pipeline generates hints for AUGUSTUS training and predicts protein coding genes, by aligning protein sequence evidence from distant organisms on the genome to annotate. The proteins used by ProtHint are the ones of the OrthoDB database, which is a combination of https://v100.orthodb.org/download/odb10_plants_fasta.tar.gz and proteins from seven species (*Anthoceros agrestis* cv. BONN, *Anthoceros agrestis* cv. OXF, *Anthoceros punctatus* (F.-W. Li et al., 2020), *Ceratodon purpureus* strain R40 (NCBI GCA_014871385.1), *Marchantia paleacea* (Rich et al., 2021), *Marchantia polymorpha* ssp. *ruderalis* TAK1 (https://marchantia.info/download/MpTak_v6.1/) *Physcomitrium patens* (Lang et al., 2018) and *Sphagnum fallax* (Healey et al., 2023)). Only the predicted genes partly supported by this protein database were kept using the https://github.com/Gaius-Augustus/BRAKER/blob/report/scripts/predictionAnalysis/selectSupportedSubsets.py script. This gave another round of cleaning and should enable to mostly predict genes present in plants, and not in contaminated sequences.

The long reads genomes of the *M. polymorpha* ssp. *ruderalis* (Bowman et al., 2017a), *M. polymorpha* ssp*. polymorpha, M. polymorpha* ssp*. montivagans* (Linde et al., 2020) and *M. paleacea* (Rich et al., 2021) were reannotated in the same fashion, as well as the two new long read genomes from the accessions CA and BoGa.

For the long reads genomes, the number of proteins predicted was comprised between 15 497 (for *M. paleacea*) and 21 625 (for our new prediction on the TAK1 reference genome). The short reads displayed way more variability, the number of proteins predicted being comprised between 11 283 (Nor-D) and 25 675 (NILSC9), with a mean value of 19 702 proteins.

The completeness of the predictions in each accession was assessed with BUSCO 5.4.4 **(Manni et al., 2021)** against the Viridiplantae odb10. The BUSCO completeness values were ranging from 95% for the long reads genomes, to 30% for low quality short reads genomes. On the average, completeness over the 135 annotated genomes was of 88.5%.

Evaluation of genes presence absence variation through orthogroup inference

The presence absence variation of the predicted proteins in the short and long reads assemblies of the accessions was determined based on orthogroup inference. The Orthofinder software **(Emms & Kelly, 2019)** was run independently on the dataset of 104 genomes for the *ssp ruderalis* pangenome, and on the dataset of 134 genomes for the Marchantia complex pangenome (the *M. polymorpha* ssp. *ruderalis* accessions RES38 and RES40 were discarded because of low BUSCO completeness, and the two annotations of the TAK1 genome, *i.e.* the reference annotation and our annotation, were taken into account). The Orthofinder analyses were run a first time with default parameters, and the resulting species trees were re-rooted (on *Marchantia paleacea* for the Marchantia complex pangenome, and on the *M. polymorpha* ssp. *montivagans* reference genome for the *M. polymorpha* ssp. *ruderalis* pangenome, following the topology of the divergence between the three subspecies found by Linde and collaborators **(Linde et al., 2020)**. Then a second Orthofinder run, forced on this tree topology, was performed, with the multiple sequence alignment parameter, to reconstruct the orthogroups using alignment and phylogenetic reconstruction methods. The resulting phylogenetic tree of all *M. polymorpha* accessions and *M. paleacea* is available in Appendix F. To remove any remaining contamination, all the proteins used in this orthogroup inference were blasted against the non-redundant protein database from the NCBI (nr version from 2023-02-20) with diamond v2.0.8 with an e-value of 1e-3, a maximum of 10 target sequences and a block size of 4. HOG with 75% of Viridiplantae diamond hits over all genes were considered as reliable, the other ones were discarded. This cleaning led to 25 648 HOG in the Marchantia complex pangenome (33 418 HOG before this final cleaning), and 28 143 HOG in the *ssp ruderalis* pangenome (initially 35 004 HOG before cleaning). Two accessions were removed from the ensuing analysis because they had less than 10 000 genes assigned to HOG (Ber-A and Nor-D).

The global pangenome therefore encompasses one genome from *M. paleacea*, 17 genomes from *M. polymorpha* ssp. *montivagans* (16 Illumina genomes + a long-reads genome), 14

genomes from *M. polymorpha* ssp. *polymorpha* (13 Illumina genomes since NILSC10 was discarded + a long-reads genome) and 100 *M. polymorpha* ssp. *ruderalis* genomes (96 Illumina genomes since RES37, RES38, RES40, Ber-A and Nor-D were discarded and 4 long reads genomes: CA, BoGa, and the two annotations of the TAK1 reference assembly).

II) *M. polymorpha's* gene-based pangenome: the outcomes

1) Pangenome of the Marchantia complex

The hierarchical orthogroups (HOG) resulting from the processing of the proteomes from the 134 Marchantia genomes were analysed in order to determine similarities and differences between the three *M. polymorpha* subspecies and the more distant species *M. paleacea*. The list of shared orthogroups among the Marchantia complex was determined by selecting core orthogroups inside each subspecies (orthogroups shared by all accessions of a subspecies, with a tolerance for the 6 low quality accessions *ssp ruderalis*, a threshold of 11 out of 14 accessions in the *polymorpha* subspecies and of 15 out of 17 accessions in the *montivagans* subspecies) and crossing them altogether and with the list of orthogroups present in *M. paleacea* (Figure 62). The HOG specific to a subspecies are HOG that appear in a least one accession of the subspecies and are absent from the accessions of other subspecies. From the 25 648 total HOGs, 2 839 were specific to *M. polymorpha* ssp. ruderalis, 331 to *M. polymorpha* ssp. *montivagans* and 372 to *M. polymorpha* ssp *polymorpha* (Figure 61).

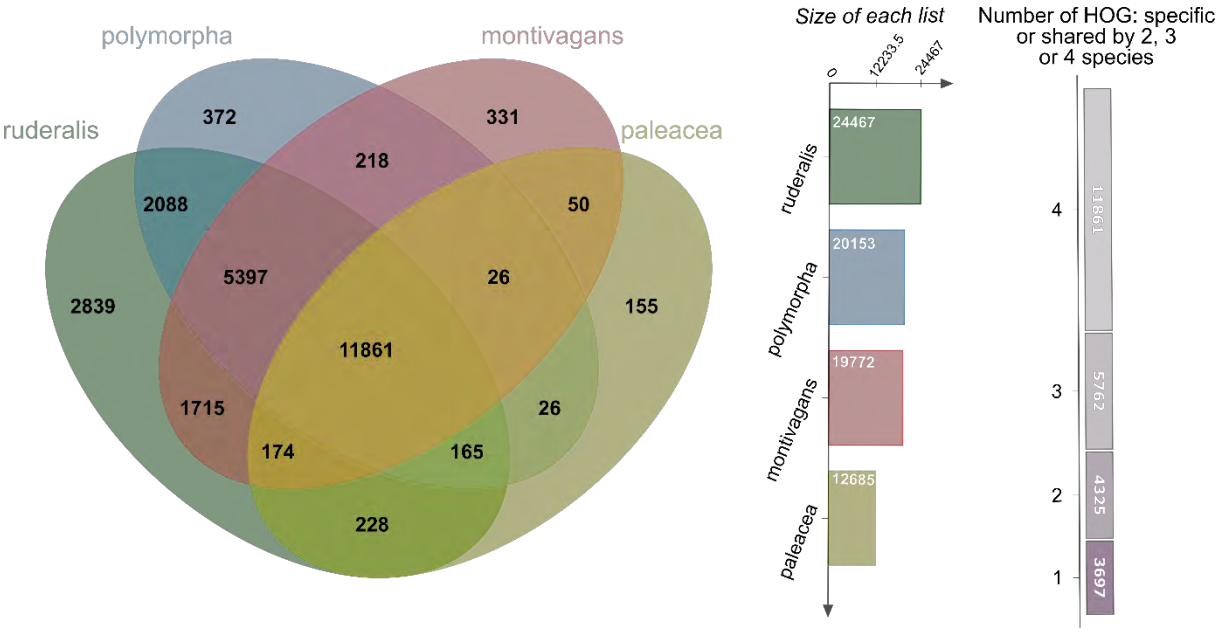


Figure 61: Overlap between the HOG present in at least one accession of the species.

Subspecies specific HOG represent genes families that evolved in each subspecies. The higher number for *M. polymorpha* ssp. *ruderalis* specific HOG is likely due to a much larger sampling in that subspecies. IPR domain enrichment have been carried on this subspecies-specific HOG lists, filtered for presence in minimum five accessions for the *ruderalis* subspecies and 2 accessions for the *montivagans* and *polymorpha* subspecies, and on the *M. paleacea* specific HOG.

The enrichment of specific domains in the *M. polymorpha* subspecies did not display very strong significance, and very diverse domains were identified (SupData3.2). Domains linked with autophagy (phagosome assembly factor), alkyl transfer to thiol (homocysteine binding domain) and retrotransposon were specifically enriched in the *ruderalis* subspecies. The *polymorpha* subspecies counted GTPase activating proteins and peptidase, when the *montivagans* subspecies had specific domains linked to mRNA processing (CID domain), or DNA repair (Mut-S protein). It is therefore hard to conclude on gene families and specific innovations linked to only one of the subspecies.

The HOG list specific to *M. paleacea* contained many domains linked to stress response: the NB-ARC and LRR domains of the NLRs, PR-4 protein, antifungal Gnk2 protein, terpenes synthases). Contrary to *M. polymorpha*, *M. paleacea* is able to engage in symbiotic interactions with arbuscular mycorrhizal fungi, and was shown to have retained central genes of the symbiotic signalling pathway shared with other land plants also able to establish this symbiosis, but lost in non-symbiotic plant lineages (Radhakrishnan et al., 2020; Rich et al., 2021). Here, the lack of identification of such specific symbiotic genes (such as: SymRK, CCamK, Cyclops) is probably due to their assignment to “orphan orthogroups” which are orthogroups containing only one gene from one species (here orthogroups with single symbiotic genes only present in *M. paleacea*), that are not displayed in the Orthofinder results.

These results show that our diversity dataset outside of the *ruderalis* subspecies may need to be enriched with other accessions to better capture the difference between the three closely related subspecies (divergence of the *M. polymorpha* subspecies around 7 MYA (Villarreal A. et al., 2016)). *M. paleacea* is more distant and its specific orthogroups are made of large gene families associated with adaptation that differentiated during *M. paleacea* and *M. polymorpha* 42 MY of separated evolution (Villarreal A. et al., 2016).

We also identified a total of 7292 HOGs shared by most accessions from all three subspecies and *Marchantia paleacea* (Figure 62). By crossing this list with the published single cell RNA-seq analysis **(Wang et al., 2023)**, we found the most significant enrichment of shared orthogroups in the expression cluster of "meristematic cells surrounding the notch" (SupData3.3 sheet 4). This cell cluster is essential for the development of Marchantia's thallus **(Solly et al., 2017)**, and contains shared genes involved in chromatin modification (MpH2B, MpH2A), cell cycle (cyclins, microtubule associated proteins, kinesin motor), or protein processing (ribosomal proteins, chaperones, proteasome) (SupData3.3 sheet 2). The shared genome of the Marchantia genus is therefore enriched in basic features of key developmental processes. However, by doing an IPR domain enrichment on the list of shared HOG, compared to all the HOG covered by the Marchantia complex pangenome, proteins related to secondary metabolism and stress response were also found, like polyphenol oxidases (production of quinones, lignin...), chalcone isomerase (production of flavonoids), or isochorismatase (salycilic acid biosynthesis). The mechanisms conserved among the Marchantia are therefore mostly related to DNA and protein processes, but a non-negligible part of stress response mechanisms is also conserved. Some of the shared genes were only described so far in the *M. polymorpha* ssp. *ruderalis* reference accession (Tak-1), like MpKARAPPO involved in gemmae development **(Hiwatashi et al., 2019)**, MpNOP1 involved in air chamber development **(Ishizaki et al., 2013)** or MpSYP12B involved in idioblast differentiation **(Kanazawa et al., 2020)**. With the study of shared genes from the Marchantia complex, these processes are extended beyond the *M. polymorpha ruderalis* subspecies.

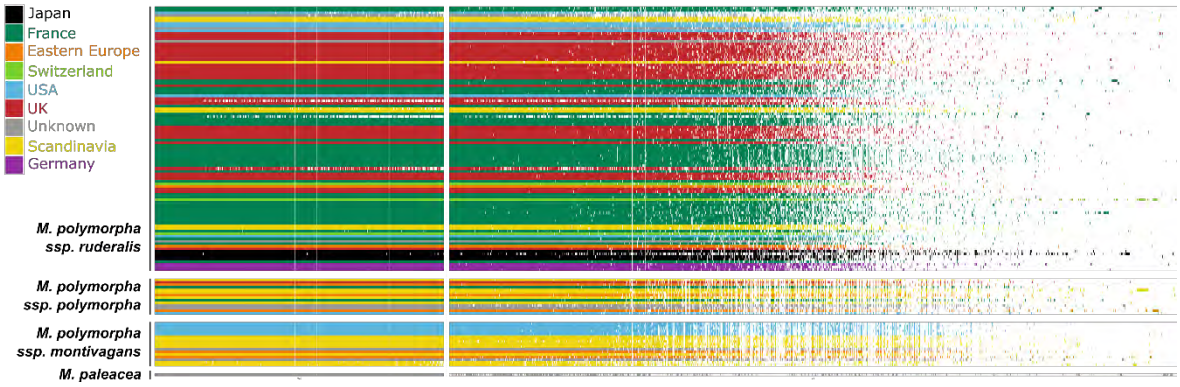


Figure 62: Landscape of gene presence/absence variation in the Marchantia genus. HOGs are sorted by their occurrence, with the genes shared by all the Marchantia species in the left part (separated from the rest by a dotted line), and the very rare genes in the right part of the matrix. The accessions (in rows) are sorted according to their position in the phylogenetic tree of the accessions generated by the software Orthofinder to infer these

HOGs, and the subspecies are separated by white lines. The colours represent the geographic origin of each accession.

2) Pangenome of *M. polymopha ssp ruderalis*

a) General description

The pangenome of the *ruderalis* susbspecies and core genome saturation was studied by determining the number of HOG present in the whole pangenome for a given number of accessions, as well as the number of HOG shared between the accessions (core genome). A list of 97 accession, with no redundancy (the long reads sequencing was kept when an accession had long and short-reads sequencing), was used to sample from two to all accessions, and count the HOG. For each given number of accessions, 500 random sampling were conducted with a custom R script. The corresponding numbers of HOG present in the sample and shared between accessions were plotted separately as boxplots (Figure 63). The number of HOG in the pangenome increases rapidly when adding the first 25 accessions, but then the augmentation slows down (with 25 accessions there is an increase of less than 0.3% of the median number of HOG covered by the pangenome with each new accession added, with 60 accessions this increase is only of 0.04%), suggesting the start of *M. polymorpha's* pangenome saturation, and therefore a closed pangenome. This type of pangenome contains a finite total number of genes/HOG when an unlimited number of accessions are added, because of the consequent number of genes shared between accessions. Modelling this saturation with a power law regression ($y = A \times x^B + C$, with y the number of HOG and x the number of accessions), as suggested by Tettelin et al. (Tettelin et al., 2005), was implemented but struggled to fit to our simulation curve. Thus, it is not possible to extrapolate on the number of accessions needed for the pangenome to fully saturate and cover all the *M. polymorpha ssp ruderalis* diversity. There is a great variability of strictly shared HOG depending on the accessions sampled, and when all accessions are taken into account, only 5500 HOG are shared by all accessions.

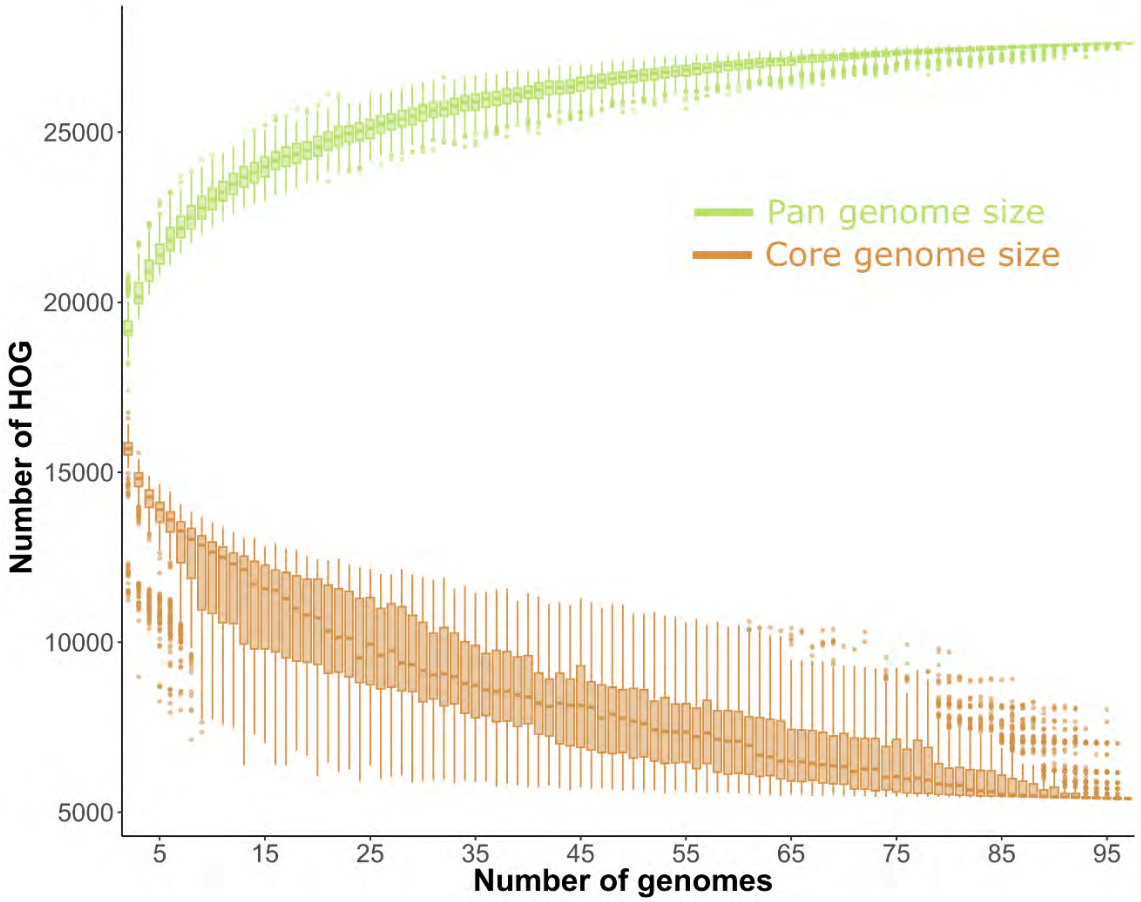


Figure 63: Modelling of the pangenome expansion (green) and of the core genome reduction (orange) when adding accessions. Our pan genome begins to saturate with 25 accessions, with a pronounced saturation around 60 accessions. Nevertheless, the number of HOG covered by the pangenome is still slightly increasing even at our maximum number of accession (increase of 0.015% of the median number of HOG between the sample for 96 and 97 accessions), suggesting that our collection almost captures the whole genetic diversity of M. polymorpha.

For the *M. polymorpha ssp ruderalis* pangenome, the core genome was determined as the HOG shared by all ruderalis accession, with a tolerance margin for 6 accessions that are very often absent from shared HOG. Some accessions are absent because of a poor sequence quality (Nor-E, RB2, RES35, NILSC19 and Mns-A), and the initial annotation TAK1 reference genome (Marpolrud_TAK1_ref_annot) is often absent from shared HOG because its annotation was not carried out with the same pipeline as the rest of the accession. These accession have a shorter orange bar (core genome) than the other accessions on the Figure 65. The accessory genome is constituted of the remaining HOG, that are present in more than four accessions, and the cloud genome is made of HOG present in four accession or less. I will refer to the core, accessory and cloud genome as “pangenomic compartments”. Since the concept of the pangenome is quite recent and at the same time increasingly used, there has been a

proliferation of terms to describe the pangenomic compartments. The part of the genome not shared by all individuals can be called dispensable, accessory, variable, or flexible. In this manuscript, I will use the term accessory genome, as I feel it reflects well the relevance of the accessory genes for each individual: even though their presence varies in the species, some of the accessory genes may be essential for a specific group of individuals to adapt to specific conditions. The same way, the genes shared by very few accessions can be called private, cloud or rare, and I will use the term cloud in the rest of the manuscript.

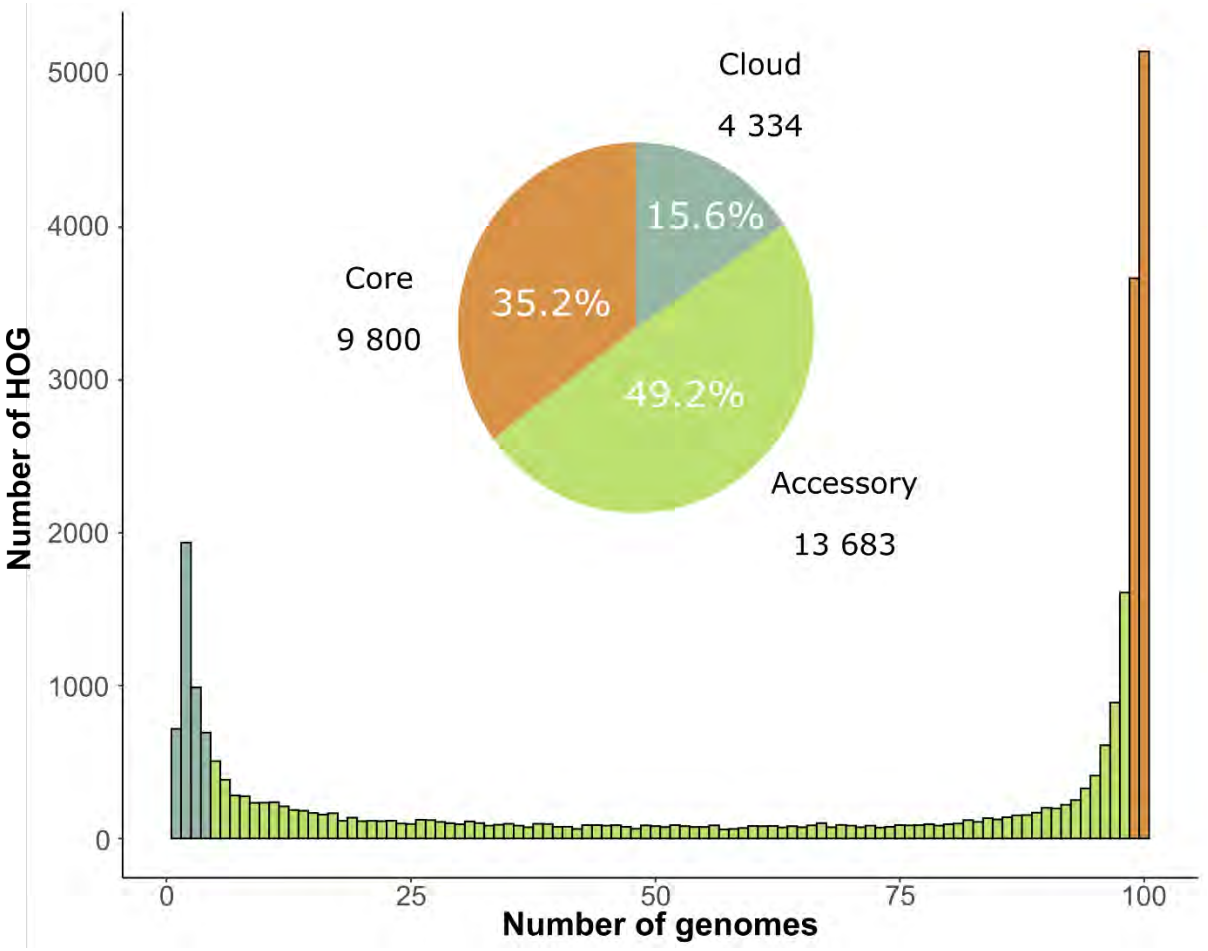


Figure 64: Distribution of the HOG depending on the number of accessions they contain and corresponding pangenomic compartments. Plot based on the HOG data of the 100 genomes from ruderalis subspecies' accessions.

The core genome, the "fundamental Marchantia's toolkit" shared by all accessions, represents 35% of the total pangenome content, whereas approximately half of the HOG belong to the accessory compartment, that contains genes only shared by some accessions (Figure 64). This flexible part of the pangenome may contain specific adaptations corresponding to local conditions, or conditions shared by a handful of accessions. The cloud genome represents a

non-negligible part of the pangenome, but it is to be taken with caution, since it is enriched in genes coming from the long read genomes (Figure 65). The genes found in this pangenomic compartment could therefore be genes present in all/most accession but not captured with the short-reads assemblies. Given this technical bias these compartments cannot be analysed as biologically rare genes. Thus, our study will focus more on the core and accessory genomes. The proportion of accessory genome varies a lot in plants (from 13.4% of genes in pigeon pea **(J. Zhao et al., 2020)** to 63% of genes in sorghum **(Tao et al., 2021)**), but the numbers found here are consistent with what has been found in other plant species, like sesame (41.79% of HOG are accessory) **(Yu et al., 2019)**, or soybean (49.9% of genes are accessory) **(Y. Liu et al., 2020)**. Nevertheless, no categorical conclusion should be drawn based on the comparison of these core and accessory genome proportions to data from other plant pangenomes, since they vary greatly depending on the sample size of accessions, on the genomes' quality, on the sequencing technology and on how the compartments were determined.

The selection signatures in core and accessory genes were compared by a T-test on Tajima's *D* statistical estimator calculated on the TAK1 reference genome. It appeared that the selection signature on accessory genes is significantly more balanced than on core genes (p-value 7.31×10^{-5}).

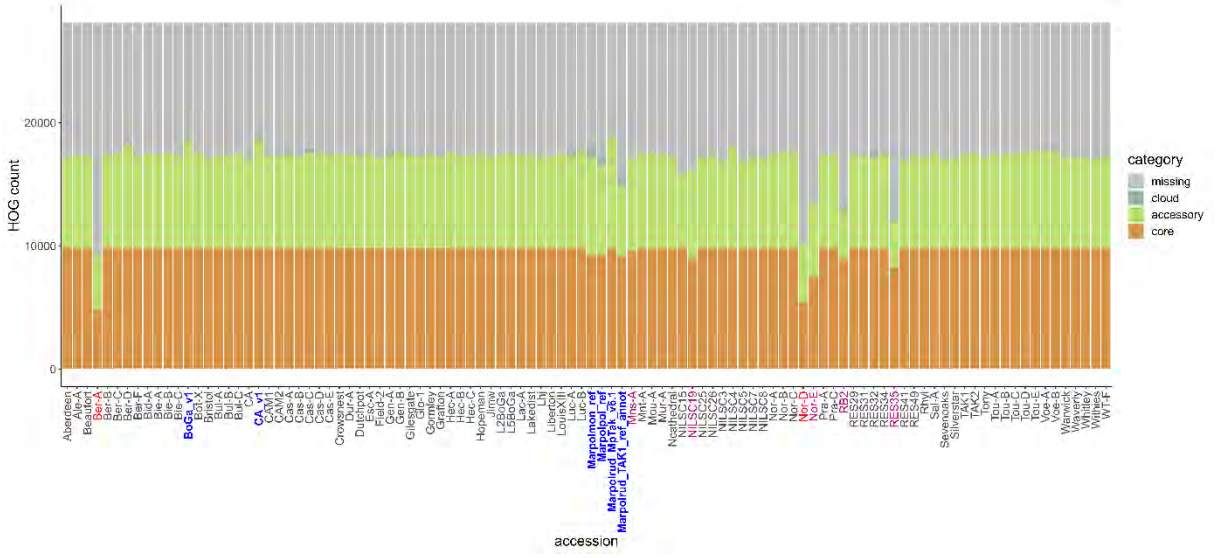


Figure 65: Assignation of the proteomes of each accession to the pangenomic compartments. This shows well the lower quality of some accessions that resulted in poor protein prediction and assignation by orthofinder. The accessions with names colored in red were discarded from the analysis, and the ones with names colored in pink were not taken into account for core HOG determination. The long reads genomes (names colored in deep blue) have more cloud genes than other genomes.

b) *M. polymorpha ssp ruderalis* core genome

IPR domain enrichment was performed on the core HOG of the *M. polymorpha ssp ruderalis* pangenome by comparing the number of occurrences for each IPR in the core orthogroups and in all the orthogroups. The presence of multiple genes in the same HOG was taken into account by multiplying the occurrence of the IPR by the median number of gene by accession in the HOG. A lot of domains were found significantly enriched in the core genome (155 before the manual curation) and many domains had very broad annotation that made it complicated to pinpoint a specific role for the genes bearing them. Nevertheless, it appears that the core genome is enriched in domains involved in transcription (SANT domain: chromatin remodelling, FCP1 domain: modification of RNA polymerase II), protein processing (TCP-1 like chaperonin intermediate domain, chaperone J domain, E3 ubiquitin ligases, protein kinases) that can enable signalisation pathways, and a variety of processes linked to DNA and RNA (DNA methylases, helicases, CTLH (microtubules chromosome segregation), elongation factor EFTu) (Figure 66, for full table see SupData3.4 sheet 2). These domains are involved in housekeeping functions and most of them are conserved in organisms, which is consistent with the same functions found to be linked to the core genome in other pangenomic studies (H. Li et al., 2022; Y. Liu et al., 2020; Lofgren et al., 2022).

Other core functions were less general and more specifically linked to plant processes, like the pheophorbide a oxygenase domain that regulates chlorophyll catabolism, the oxoglutarate iron dependent dioxygenase that catalyse formation of phytohormones (ethylene and gibberellines), the EXORDIUM like domain that mediates brassinosteroids promoted growth, lipoxygenases that acts in lipids oxydoreduction and can lead to the production of jasmonic acid, or the recently characterised PADRE domain (pathogen and abiotic stress response, cadmium tolerance, disordered region-containing) that responds to various biotic and abiotic stresses (Didelon et al., 2020).

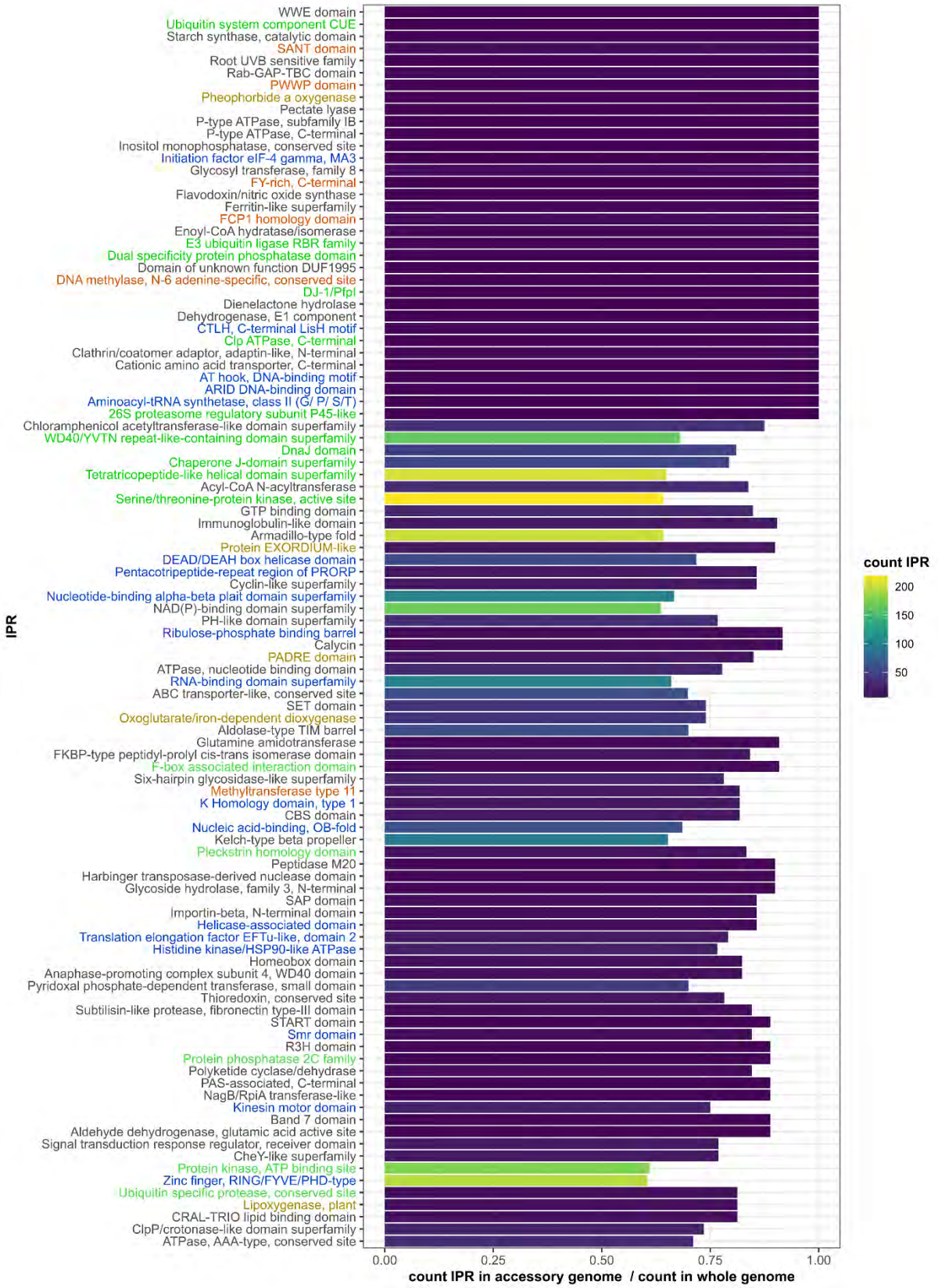


Figure 66: Selection of enriched IPR domains (FDR q-value <0.05) in the core genome of M. polymorpha ssp ruderalis. The list of significant IPR domain initially counted 154 terms, but was manually curated to remove redundant terms that designated the same orthogroups or the same function, and is ordered in ascending order of q-value. IPR domains linked with the aforementioned categories had their names coloured: green are linked with

proteins processes, blue with DNA/RNA processes, rust with transcription regulation and light brown with plant specific mechanisms.

When crossing the lists of core and accessory genes with scRNA-Seq analysis, we found that the expression cluster corresponding to "cells surrounding the notch" was the most significantly enriched in core *M. polymorpha* ssp. *ruderalis* genes, similarly to what was found for the Marchantia pan-genome. Other cell categories with more core genes expressed were air pore cells and rhizoids. These synapomorphies specific to some (or all) bryophytes, are central organs of the plant, governed by conserved genes.

c) *M. polymorpha* ssp. *ruderalis* accessory genome

The IPR domain enrichment of the accessory genome was performed in the same fashion, leading to a list of 140 significant IPR terms. The main functions of these IPR terms are detailed on Figure 67 (for full table see SupData3.4 sheet 3). Most domain found here can be linked to plant response to various stresses. For example, the LRR and NB-ARC domains from the NLR proteins are significantly enriched, showing the role of the accessory genome in pathogen response. Corroborating that, the LURP-1 related protein that is involved in the basal defence of *A. thaliana* against pathogens **(Baig, 2018)** is also enriched, as well as the allergen V5 domain, characteristic of PR1 (pathogenesis related class one) proteins. PR proteins are core components of the inducible innate defence of the plant during biotic stress but also abiotic stresses. The PR1 class is the most strongly expressed type of PR protein, whose exact biochemical activity is still unclear. It displays antimicrobial activity, and induces general stress resistance in the plant, probably by activation of a signalling pathway and interaction with other PR proteins **(Han et al., 2023)**. Some domains more characteristic of response to abiotic stresses were also found, like the aquaporin membrane channels that allow the transport of water but also gases and some nutrients and enhances plant tolerance to various abiotic stresses like dehydration, salinity and heat **(Maurel et al., 2015)**. Similarly, there is an enrichment of the ABA-WDS domain typical of the ASR (ABA-stress-ripening) protein family that function as both chaperones and transcription factor and are involved in response to abiotic stresses. Other general stress response candidates appear in the enriched domains like the peroxidases that play a part in stress response through diverse pathways (modification of the plant cell wall, metabolism of Reactive Oxygen Species (ROS) involved in signalling, synthesis of the antimicrobial phytoalexins…) **(Almagro et al., 2009)**. The accessory genome is also enriched in chalcone synthases, key enzymes of the flavonoid pathway that enables the accumulation of

antimicrobial (phytoalexins), insecticidal, antioxidant and UV-protecting components (Dao et al., 2011). Finally, polyphenol oxydases can also be cited, since they are induced in response to biotic and abiotic stresses an produce reactive molecules (o-quinones) that have toxic effects on herbivores, pathogens and enable other defences in the plant like physical barriers (J. Zhang & Sun, 2021). More surprisingly, the ribonuclease H and DNA/RNA polymerase superfamilies are also represented in the pangenome, which may be due to the large size of both gene families, that enables variation to occur inside of the superfamilies.

Many of the domain coincide with functions that were found significantly associated with response to environmental conditions or to the fungal pathogen *Colletotrichum nymphaea* in the genome-wide association studies. Peroxidases, heavy metal associated domain, NLRs, LURP-1 related proteins, Lateral organ boundary transcription factor and cupins were found in both analyses. The isoprenoid synthase domain was enriched as well, re affirming the key role of terpenes in *M. polymorpha*'s response to stresses. The accessory genome of *M. polymorpha* is therefore clearly associated with its adaptation to various stressors.

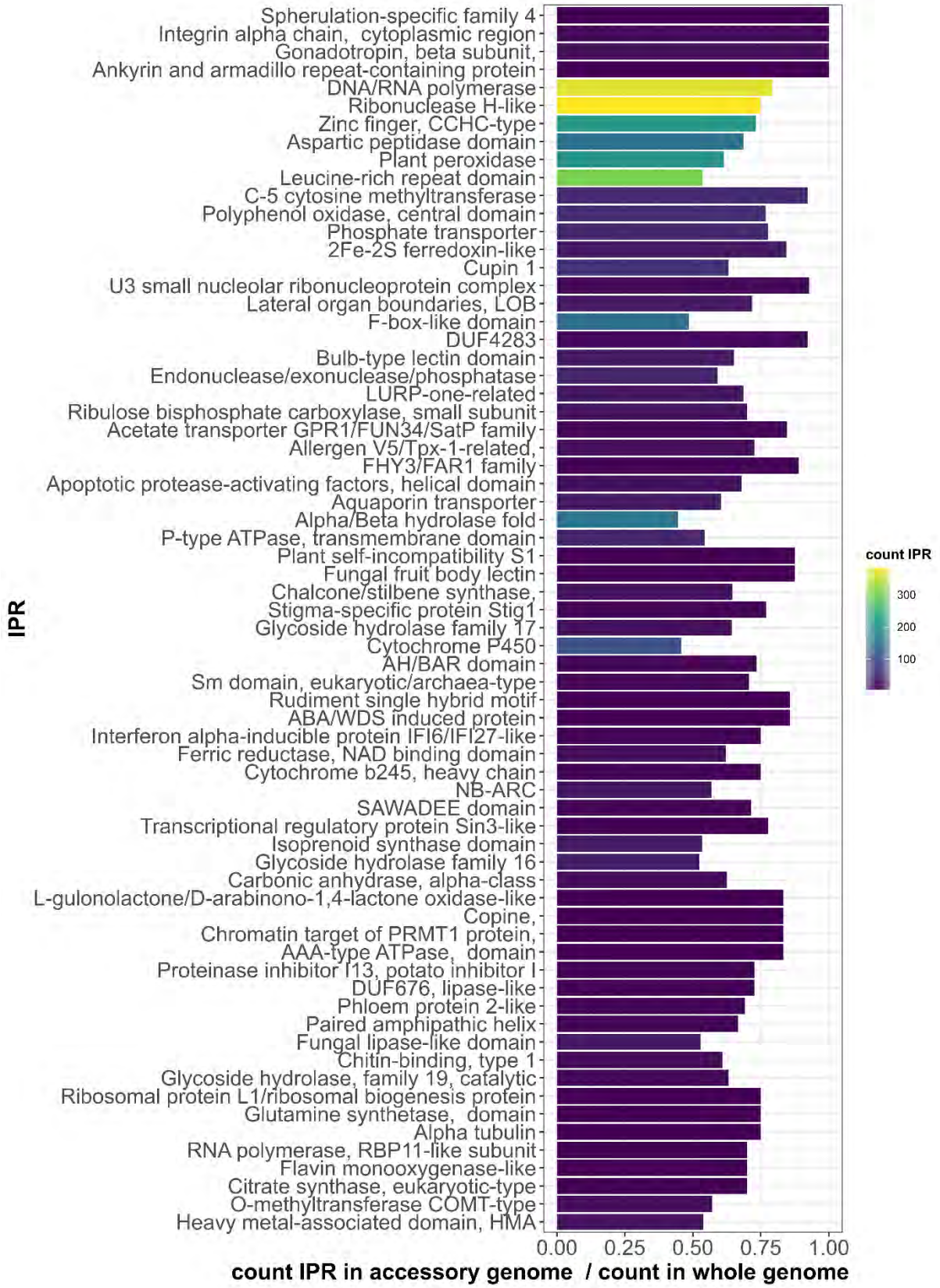


Figure 67: Selection of enriched IPR domains (FDR q-value <0.05) in the accessory genome of M. polymorpha ssp ruderalis. The list of significant IPR domain initially counted 139 terms, but was manually curated to remove redundant terms that designated the same orthogroups or the same function, and ordered in ascending order of q-value.

A lot of the families described here are classical plant genes already known to be involved in stress response in most land plant clades. But one of the candidates from the enrichment departed from this tendency: the fungal fruit body lectin, that seemed absent from most land plants. Lectins are a large class of carbohydrate binding proteins found in all kingdoms of life. In plants there are 12 different families of lectin that have various evolutionary origins and biological activities (Van Holle & Van Damme, 2019). The fungal fruit body lectins, that can also be referred as *Agaricus bisporus* agglutinins, have been discovered in *Marchantia polymorpha* in 2007 (Peumans et al., 2007). They have arisen through horizontal gene transfer (HGT) originating from the fungal ancestor of the Basidiomycota and Ascomycota clades, and have been found so far in some liverworts and some mosses (Ma et al., 2022; Peumans et al., 2007; Van Holle & Van Damme, 2019). To verify its distribution in the land plants, we searched for orthologs of the 9 fungal fruit body lectins present in the reference genome of *M. polymorpha ssp ruderalis*, in a database containing genomes from all the main clades of landplants (SupData2.4). It was found in liverworts, in mosses, but also more surprisingly in genomes from *Adiantum nelumboides* and and *Alsophila spinulosa* and in 20 other monilophyte species' transcriptomes. In previous studies on plant lectins, these Agaricus bisporus agglutinins were only found in mosses and liverworts but the presence of these lectins in ferns has probably been missed because the only genomes available were from aquatic ferns that may have lost these genes with their return in water (Baggs et al., 2020).This presence of fungal fruit body lectins in monilophytes gives a whole other perspective on the gain of these domains in plants: they might have been transferred from a fungus to the ancestor of land plants and then lost in multiple clades (Figure 68, and see Figure 6 from the introduction for comparison with land plant clade distribution). Six of the nine fungal fruit body lectin genes present in TAK1 are differentially expressed in stress conditions. Most of them are downregulated when the plant is subjected to dark, nitrogen deficiency or *P. palmivora* infection, and up regulated in drought conditions induced by mannitol application.

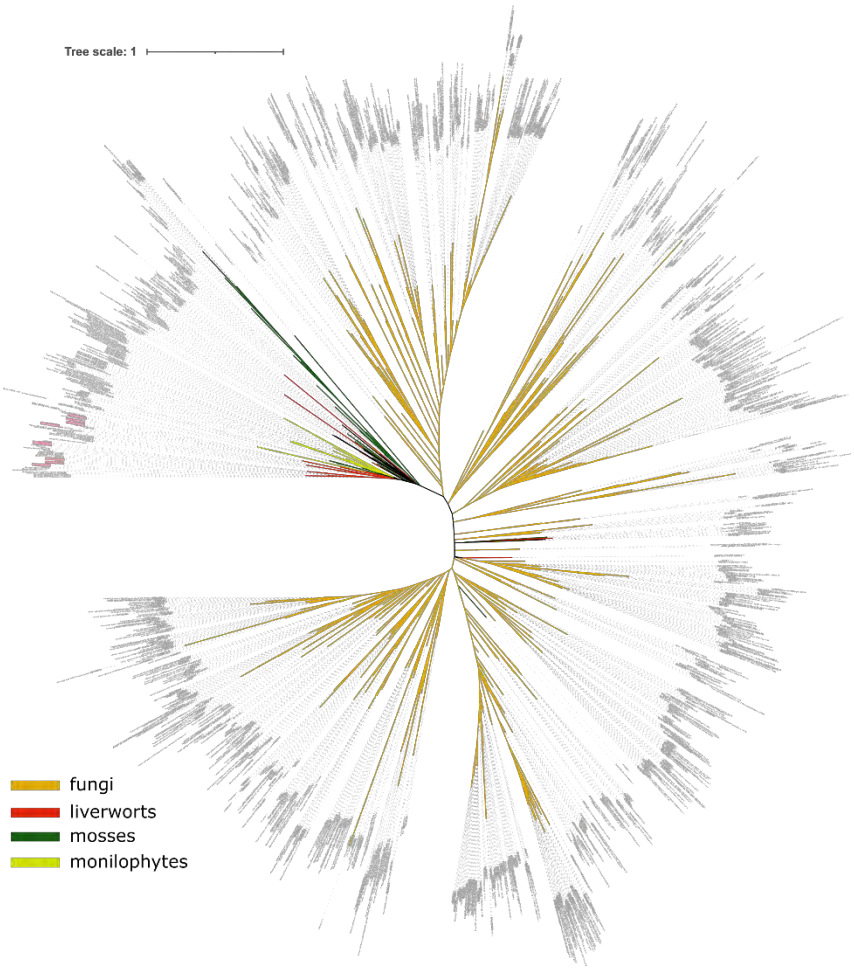


Figure 68: Phylogenetic tree of the orthologs of the fungal lectin genes present in the reference genome of M. polymorpha ssp ruderalis. Most orthologs are fungal ones (orange), clearly showing the fungal origin of the genes. In land plants, the genes are present in liverworts, mosses but also monilophytes.

When crossing the list of accessory genes with the scRNAseq analysis **(Wang et al., 2023)**, it appeared that the expression clusters with a significant proportion of accessory genes were the ones linked to *gemma cells* and *oil body* cells. Among the 283 genes of the gemma cell cluster, 83 are accessory and are mainly linked to secondary metabolism, such as chalcone synthases or cytochrome P450. For the *oil body* cell cluster, 52 of the 177 assigned genes are accessory, many of which are also involved in secondary metabolism (SupData3.5). This includes some microbial terpene synthase like genes and polyphenol oxidases, probably allowing the on-site synthesis of the oil body metabolites. When comparing core and accessory genes expressed in the oil bodies and linked to terpene synthase, it is possible to observe a clear pattern between the position of the gene in the pathways and their membership in the pangenome compartments: core genes are located upstream of the pathway, like the precursor synthesizers IDS genes, whereas the accessory genes are terminal genes like MTPSL.

d) Comparison of *M. polymorpha ssp ruderalis* core and accessory genome expression

In order to understand in which conditions the core or the accessory genome are more likely to be expressed, the core and accessory genes list from the reference genome were crossed with down-regulated or up-regulated genes in each condition from the RNA-Seq studies in response to *P. palmivora* (Carella et al., 2019) and abiotic stresses (Tan et al., 2023). The proportion of accessory and core genes up-regulated and down-regulated in each RNA-Seq condition was compared to the proportion of core and accessory genes in the whole reference genome (TAK1_V6, initial annotation), using a Fisher exact test. This lead to a list of conditions with significantly (p-value <0.05) more genes from one compartment being either up-regulated or down-regulated (Figure 69).

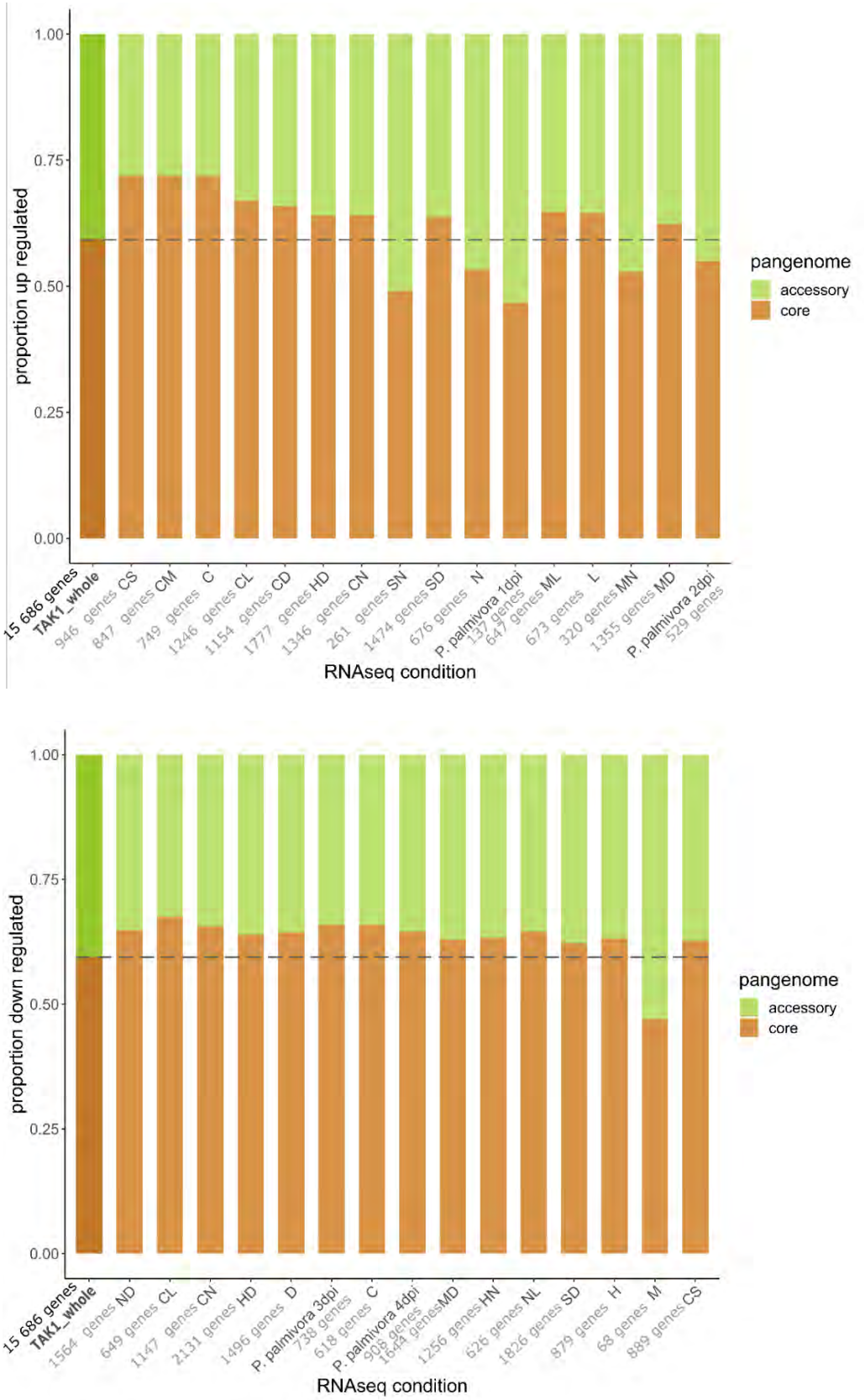


Figure 69: RNAseq conditions for which the core or accessory genome are significantly more up regulated (upper part) or down regulated (lower part). Only the RNA-Seq conditions with a significant difference between deregulation in a compartment and deregulation in the whole genome are represented. The first bar represents the proportion of core and accessory genes in the whole reference genome on which the RNA analysis were carried

out. The conditions with the green bar going below this reference proportion are the ones for which the accessory genome is more deregulated. Abbreviations of the abiotic stress conditions are the following: N stands for nitrogen deficiency, D for dark, C for cold, H for heat, M for mannitol, L for light, S for salt. The stresses can also be cumulated, hence the letters associations.

It appears that in general, the deregulated genes are mostly from the core genome, which is confirmed by the Fisher exact test ran on all the differentially expressed genes (DEG) from the reference genome: 5411 core genes out of 9322 (58%) are differentially expressed in at least one condition, versus 3106 accessory genes DEG out of the total 6364 genes (49%), which gives a p-value of $2.2x10^{-16}$ in favour of core genome deregulation. For downregulation, only the mannitol condition (osmotic stress) seems to recruit more accessory genes, but it is a condition with a low number of genes differentially expressed. For upregulation, five conditions display more accessory genes deregulated: three are linked to nitrogen deficiency (alone, or in combination with salt stress or mannitol stress), and two are linked to early stages of infection by *P. palmivora* (1 and 2 dpi). The accessory genes up-regulated in these conditions are typical from the accessory genome: LURP-1 protein, ABA/WDS induced protein, Aquaporins, PR-1 proteins, chitinases, germins, and many peroxidases. The nitrogen deficiency is the only nutrient deficiency tested in the RNA-Seq experiment. A quite speculative hypothesis could be that the accessory genome differential expression is important in biotic interaction conditions. This biotic interaction could be either pathogenic interactions (in the case of the *P. palmivora* infection) or communication with the soil microbiota in order to optimise nutrient uptake **(Hartman & Tringe, 2019)**. Overall, this comparison indicates that the core genome of *M. polymorpha ssp ruderalis* is more targeted by deregulation than its accessory genes, but a non-negligible part of the accessory genome is transcriptionally triggered, especially in some stress conditions.

4) Presence-absence-based genome-wide association studies

In plants, structural variations, such as presence-absence polymorphisms, account for a non-negligible part of the intra-specific genome variability **(Saxena et al., 2014; Yuan et al., 2021)**. Studies have shown that large structural variations, like insertions-deletions (InDels), gene presence-absence variations (PAV) or copy number variations (CNV) can play important roles in the genetic bases of phenotypic variation in angiosperms **(C. Liu et al., 2022; J.-M. Song et al., 2020; Y. Sun et al., 2022)**. Moreover, genomic regions missing from the reference genome can be overlooked when working with SNPs called on this genome, although they can bear

genetic polymorphisms linked with phenotypic variation. Therefore, in order to complete our GWAS and GEA results obtained on reference-based SNPs, we used genome wide association studies to cross the presence-absence variation (PAV) detected in the accessions with the pangenome analysis, with either the climatic data from Marchantia's sampling sites or the phenotypic response of *M. polymorpha* to the pathogenic fungus *C. nymphaeae.* To do so, we used the presence-absence polymorphism of the HOG in the ssp. *ruderalis* pangenome and recoded it as 0 if the orthogroup is absent in an accession and 1 if it is present, regardless of the number of genes it contains. For the GWAS on the pathogenic data, these presence-absences variations were compared with phenotypic data from 69 accessions (with a specific focus on the direct response to the pathogen: brown area and brown percentage phenotypes). For the GEA analysis, two PCA analysis were conducted on the climatic data from 90 accessions, one on the bioclimatic variables linked to temperature (BIO1 to BIO11) and the other on the bioclimatic variables linked to precipitations (BIO12 to BIO19), similarly to what has been done for the SNP-based analyses. The genome wide association study was then focused on the first two principal components of each PCA (that account for 87.91% of the variability due to precipitation-linked variables and 87.05% of the variability due to temperature-linked variables), to reduce the number of variables to consider for this PAV-GWAS trial. Some of the 17 806 presence-absence variations present in the 90 accessions had the same pattern of presence absence (they were present in the exact same accessions). In this case only one HOG with this PAV pattern was kept to represent the others and have less sites to test. The 12 312 resulting PAV sites were tested against the climatic PCA dataset with a linear mixed model from the GEMMA software (options `-lmm 4 -maf 0 -miss 1`), corrected for population structure by a kinship based on the SNPs from the 90 accessions. For the GWAS based on the brown area and brown percentage of the thallus, the 16 796 PAV HOG were also de-duplicated to keep only one HOG with a given PAV distribution in the accessions, leading to 11 693 PAV that were used in the GEMMA software, with the same parameters as in the PAV-GEA analysis, and a correction for population structure brought by a kinship based on the SNPs present in the 69 accessions. The results were then filtered to keep only PAV HOG with a MAF of 0.05, equivalent to a polymorphism present in at least 4 accessions for the GWAS data (10 754 PAV HOG kept) and 5 accessions for the GEA data (10 902 PAV HOG kept). Since HOG from the pangenome could not be anchored on the chromosomes, the results were represented as QQplots, like the ones on the thallus browning phenotypes displayed Figure 70.

a) PAV-GWAS analysis on *M. polymorpha* symptoms of *C. nymphaeae* infection

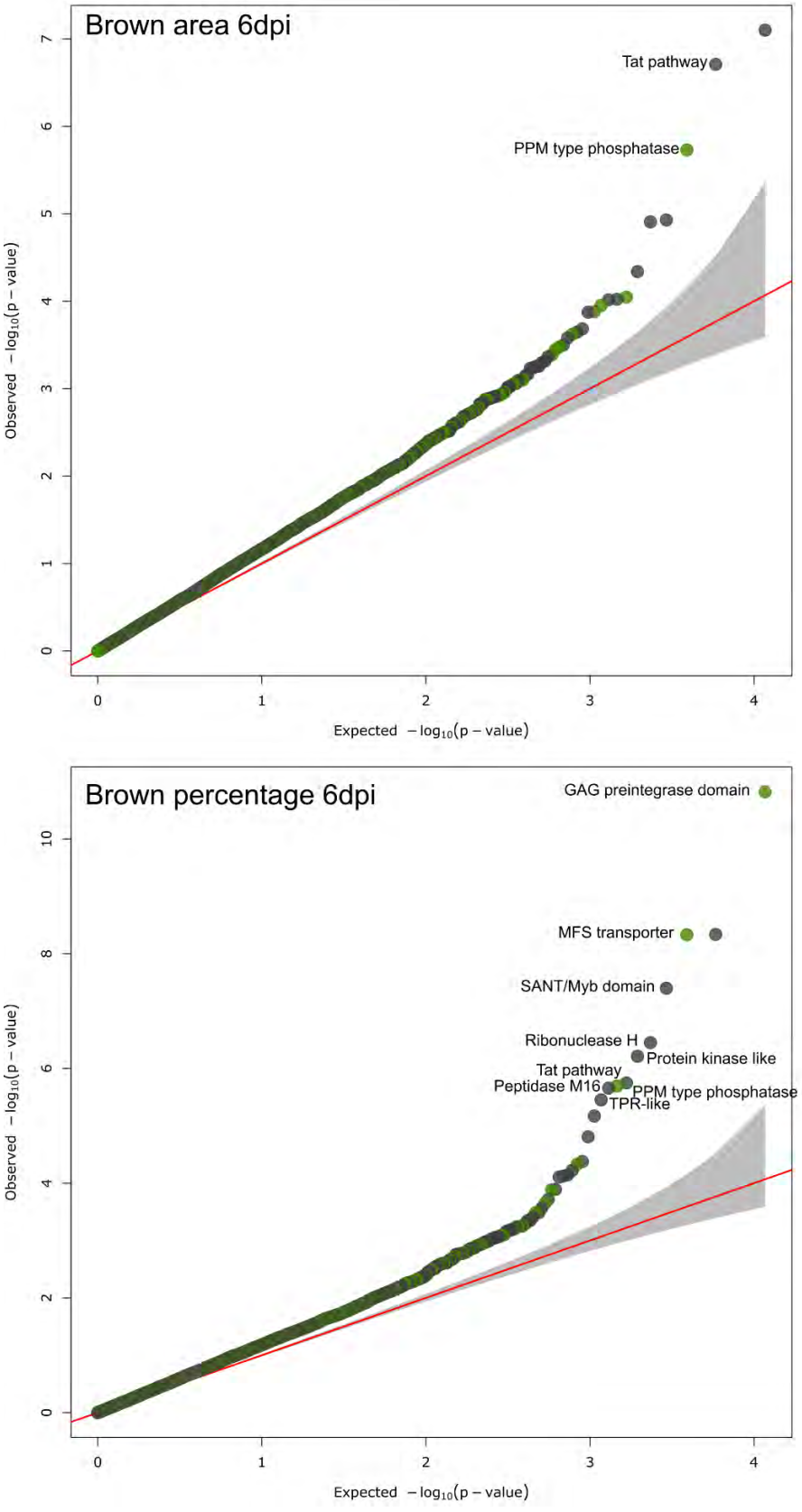


Figure 70: Expected and observed p-values (QQplot) of the association of HOG displaying PAV in the 70 accessions in the ssp. ruderalis pangenome with the variation of response to C. nymphaeae infection. Each dot represents an HOG (or group of HOG with the same PAV pattern in the accessions), and the green dots symbolise the HOG not

present in the reference genome. The functional annotation of the HOG with p-values below 4.6x10^{-6} is specified when they have one.

I considered all the PAV-GWAS candidate HOG with a p-value below 4.6x10^{-6} (Bonferroni correction: 0.05/10 754, full list in SupData3.6), and visualised their PAV pattern in the phenotyped accessions (Figure 71). Some candidates' absence seems to be linked with susceptibility to *C. nymphaeae*, like the twin arginine pathway proteins that has a role in protein transport across membranes in organelles (mitochondria and chloroplasts), or the Myb-like DNA binding domains that could be involved in gene expression regulation. On the contrary, the presence of some genes in the accessions' genome seems to correlate with an increase of the symptoms, like the protein kinase or the tetratricopeptide repeat (TPR) family. TPR proteins can be involved in a variety of cellular processes, some of which can be linked with plant response to stress. For instance the TPR-containing gene Bsr-k1 confers broad spectrum resistance to rice, against bacterial and fungal pathogens (X. Zhou et al., 2018).

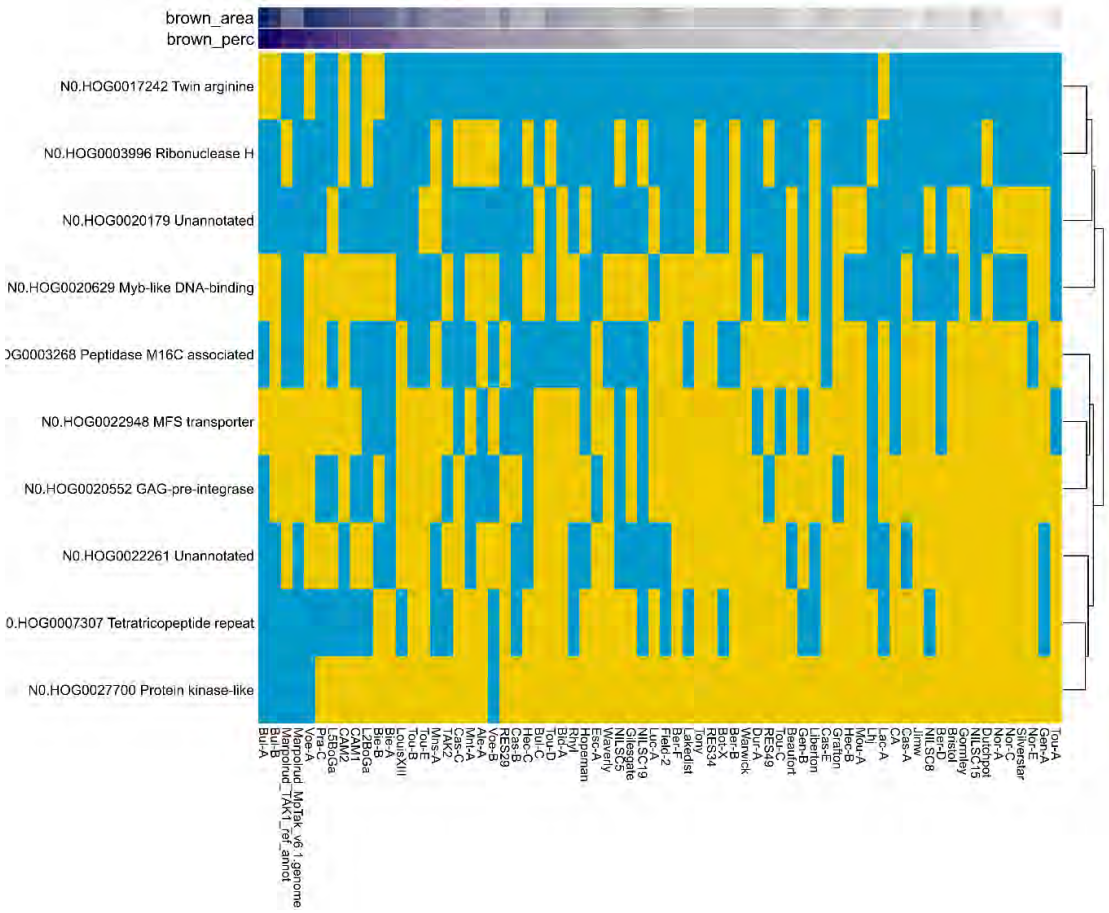


Figure 71: Presence (in blue) or absence (in yellow) variation of the different candidate HOG in the phenotyped accessions. The PAV matrix is clustered according to the browning phenotypes of the 70 accessions, represented on top of the matrix.

b) PAV-GEA analysis on climatic data

The results of the genome wide analysis studies for the two first principal components of precipitation and temperature linked variables are represented in Figure 72 and Figure 74, with the functional annotation of the HOG with the smallest p-value. For the analysis, only HOG with a p-value lower than $4.6x10^{-6}$ (Bonferroni correction: 0.05/10 902) were considered (full tables in SupData3.7).

Precipitation-linked bioclimatic variables

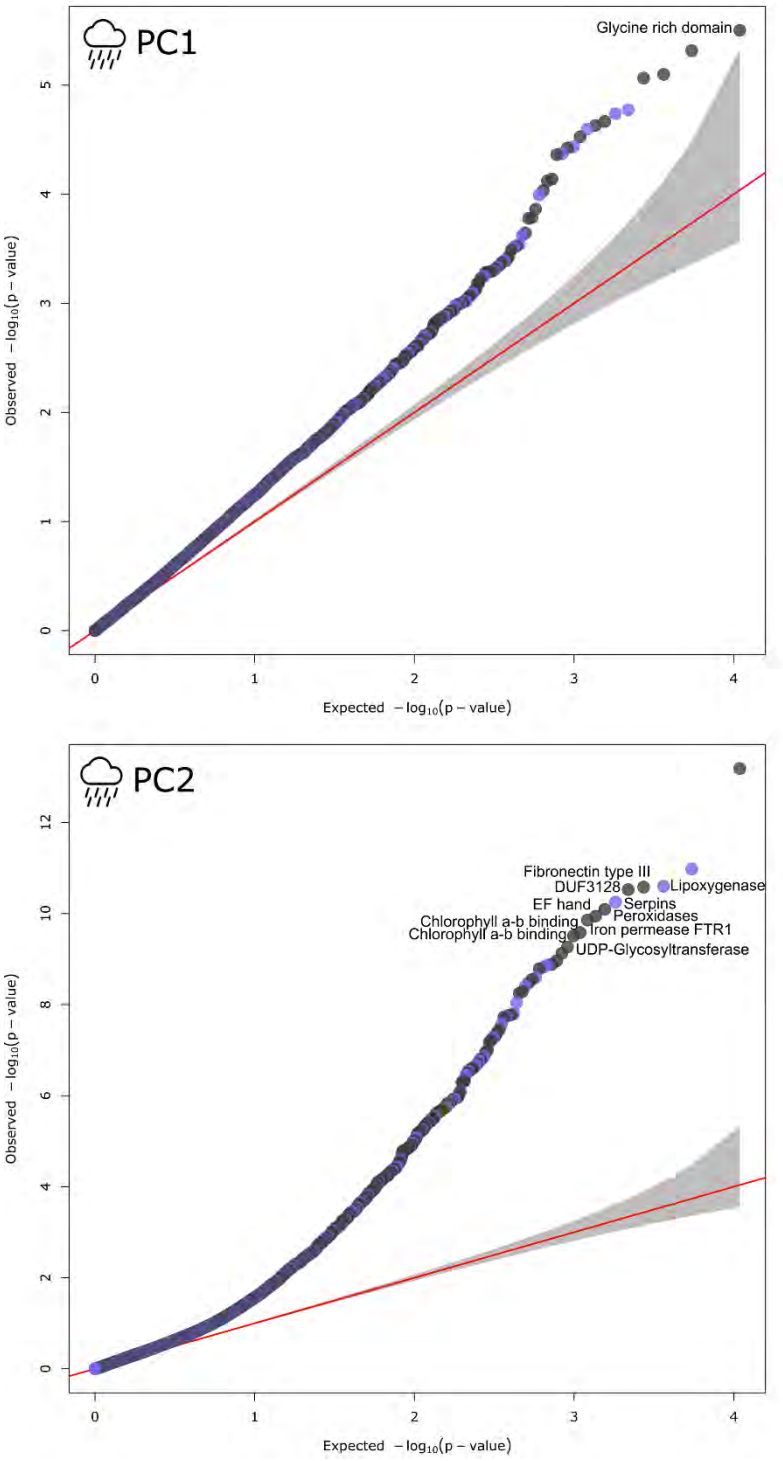


Figure 72: QQplots for the two first principal components of precipitation related variables in the 91 accessions of M. polymorpha ssp. ruderalis. Each dot represents an HOG (or group of HOG with the same PAV pattern in the accessions), and the purple dots indicate the HOG not present in the reference genome. The functional annotation of the HOG with p-values below $4.6x10^{-6}$ is specified when they have one (and when there is enough space on the plot).

For the precipitation linked variables, the PAV matrix represented alongside the PC1 and PC2 of the precipitations (and ordered according to these principal components) shows that most

significant HOG candidates mainly display a presence absence variation between the reference genome (bottom lines of the Figure 73) and the rest of the accessions. Candidate HOGs only display presence in most accessions or absence in most accessions, with the “minor allele” accession always being the reference genome. This sharp pattern may be due to the distribution of the accession on the PCA on precipitation related variables. Most accessions have really similar values of precipitation variable, except from a few accessions that have very distant values (Figure 27), among which is the reference genome (TAK1). This could cause a bias towards TAK1 in the statistical test and lead to the striking pattern shown in Figure 73. A few HOG display this pattern of absence in most accessions and presence in the reference genome, but are also present in other accession with a strong precipitation phenotype, like RES34 and or CA, which makes them stronger candidates. Among them the HOG0005554 encodes for UDP glycosyl transferase, like the Mp6g15860 gene in the reference, also known as MpUGT26. UGT modify small molecules like secondary metabolites or phytohormones by attaching them sugar molecules. They can therefore be involved in the plant’s protection against biotic and abiotic stresses (Gharabli et al., 2023). Another candidate, the HOG0001661 encodes for a peroxidase (Mp5g13830 also known as MpPOD105) which is consistent with the peroxidases found associated with the climatic variations and pathogenic responses in the genome wide association analysis on SNPs. Finally, another candidate is the HOG0025637 (containing Mp6g11250) that encodes for a protein with NB-ARC domain, that was not found among the list of known NLRs of the reference genome.

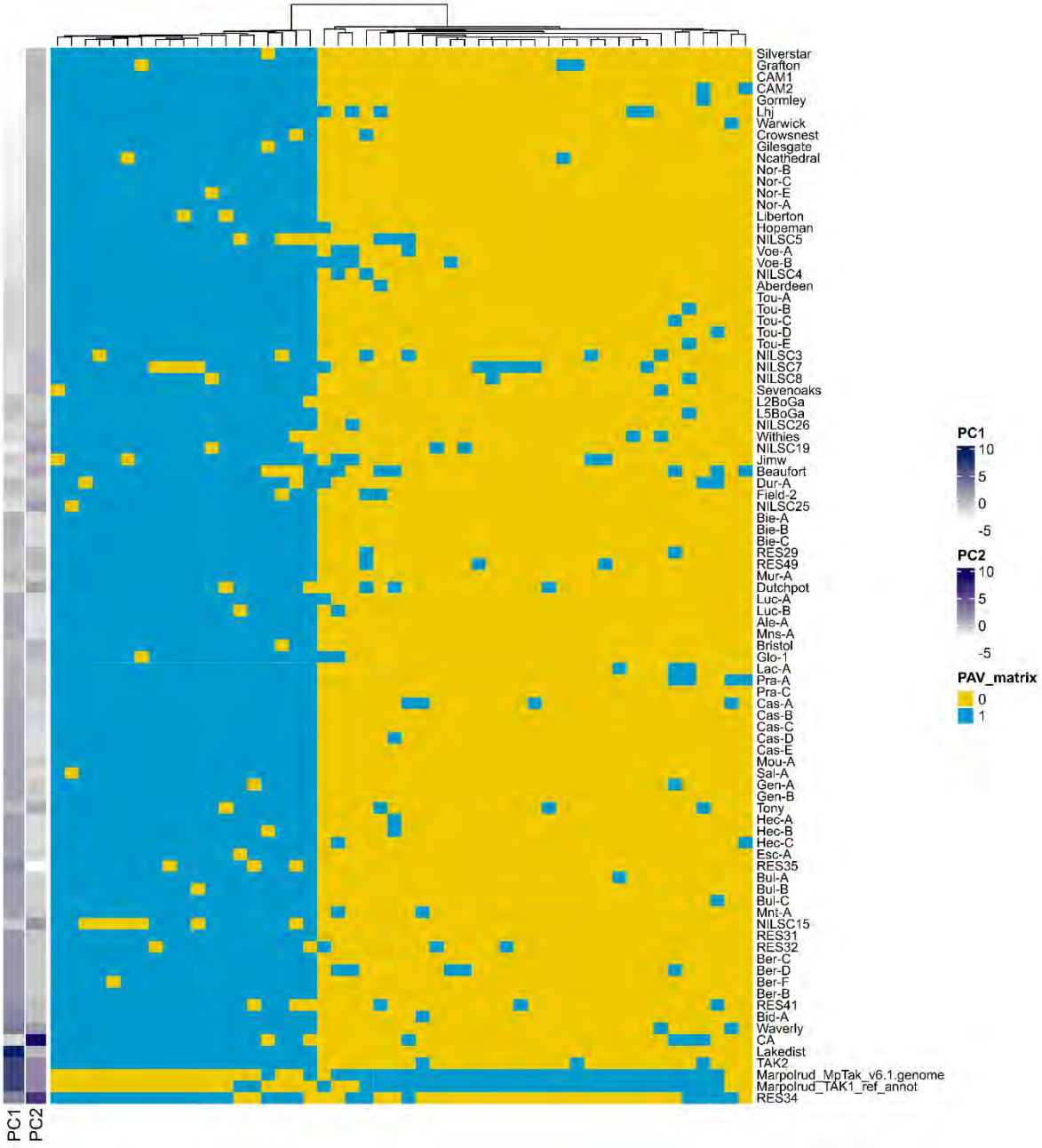


Figure 73: matrix of the presence absence variations in 90 accessions of 50 candidate HOG correlated with precipitation variation between sampling sites. The accessions are ordered by their position on the PC1 and PC2 computed on precipitations variables, and the candidate genes (HOG) are clustered according to their PAV patterns (presence in blue, absence in yellow).

Temperature-linked bioclimatic variables

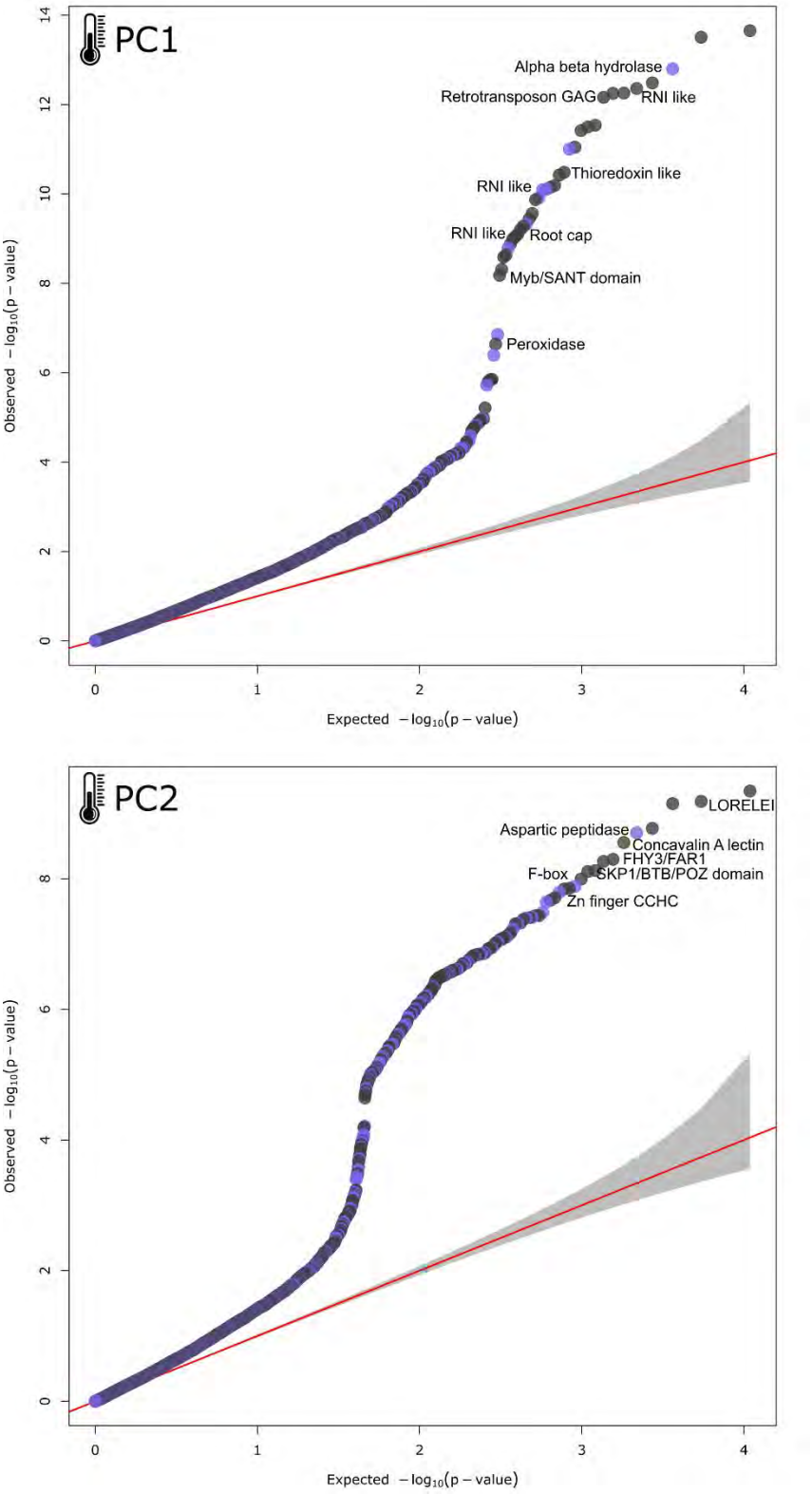


Figure 74: QQplots for the two first principal components of temperature related variables in the 91 accessions of M. polymorpha ssp. ruderalis. Each dot represents an HOG (or group of HOG with the same PAV pattern in the accessions), and the purple dots indicate the HOG not present in the reference genome. The functional annotation of the HOG with p-values below $4.6x10^{-6}$ is specified when they have one (and when there is enough space on the plot).

The same analysis of the PAV matrix was carried out on 95 candidate HOGs of the temperature dataset (Figure 75). The presence-absence patterns of the candidates are not always opposing the reference genome to the other accessions, contrary to the results obtained for the precipitation variations. indeed, there is a much more heterogeneous patterns of geographical variation for temperature variables, and TAK1 is not particularly contrasted relative to other accessions (see PCA plot for temperature variables, figure 26). Once again there is two clear groups of candidate genes: genes present in most accessions (columns with mostly blue squares) and genes absent in most accessions (column with mostly yellow squares). A lot of the candidates are reminiscent of gene families found in other analysis: cytochrome p450, lipoxygenase, leucine rich repeats, major intrinsic protein, peroxidases, thioredoxin, or even a fungal fruit body lectin (FFBL). The FFBL is only present in 10 accessions among which no long-reads genomes, it therefore needs further confirmation to be sure it is not just contamination common to these 10 accessions. Some other candidates are from "new" families, like the LORELEI HOG (Mp5g09600 also known as MpLRE1) that transduce external cues coming from CrRLK1L like FERONIA into signals inside the cell **(Mecchia et al., 2022)**. This LRE gene is duplicated in Marchantia which could explain the PAV, and also its strong signature of balancing selection (top 0.3% of genes under balancing selection). Other candidates are from the Tapetum determinant 1 like family (TPD1) that is crucial for pollen development in flowering plants but also conserved in other landplants and controls signalling of developmental pathways **(B. Zheng et al., 2019)**. There is also many RNI-like proteins, that bear a LRR domain and a COR (C-terminal of Roc) domain, that seem to be involved in the formation of ubiquitin ligase complexes (like the COI1 gene in *A. thaliana* that belongs to this family).

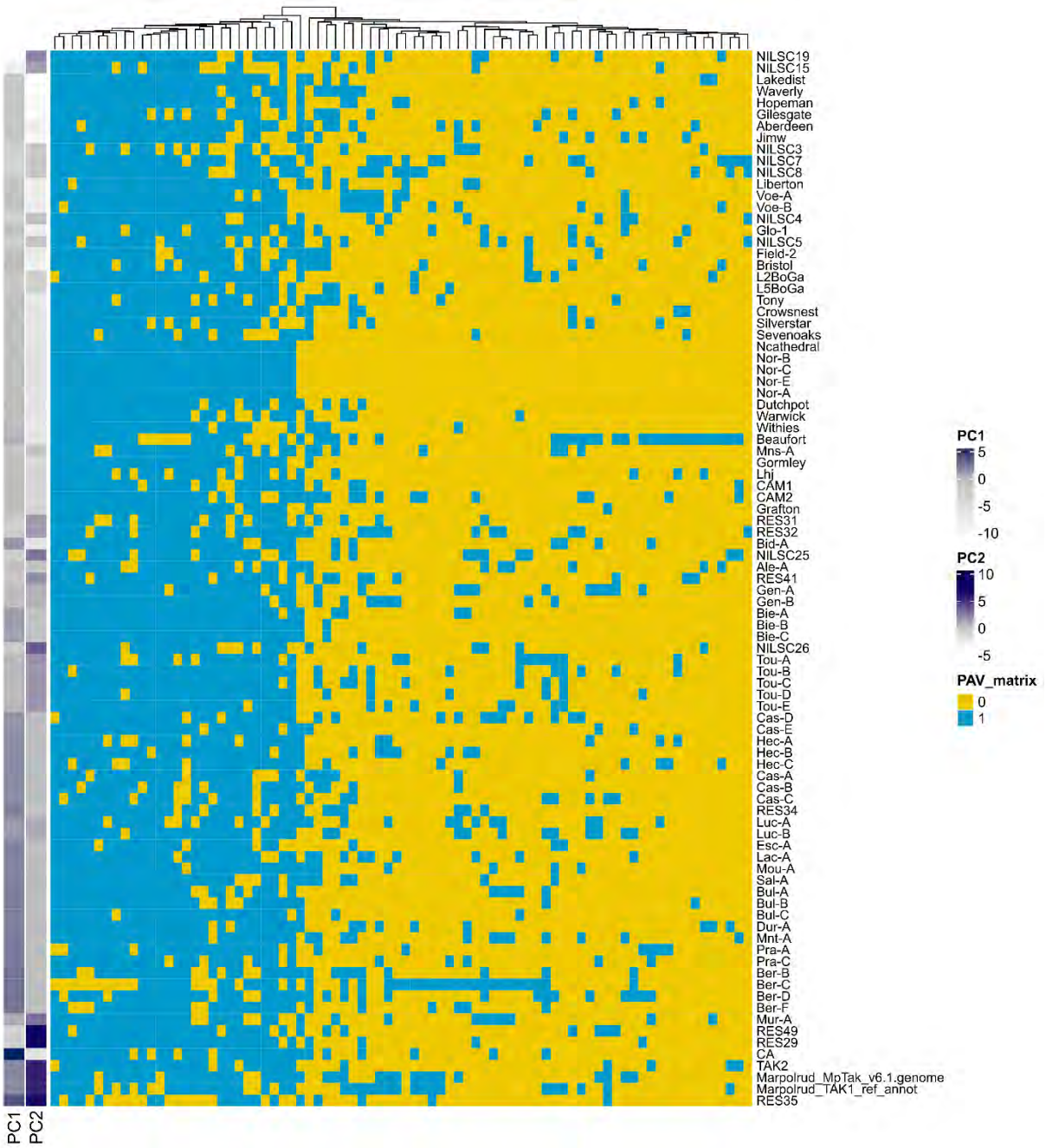


Figure 75: matrix of the presence absence variations in 90 accessions of the 91 candidate HOG correlated with temperature variation between sampling sites. The accessions are ordered by their position on the PC1 and PC2 computed on precipitations variables, and the candidate genes (HOG) are clustered according to their PAV patterns (presence in blue, absence in yellow).

These presence-absence genome-wide associations studies in *M. polymorpha* ssp. *ruderalis* have shown that a non-negligible part of the associated HOG are not present in the reference genome. In the pathogenic GWAS analysis, two of the ten significant candidates were not present in TAK1 (one being a GAG pre-integrase, but the other being a major facilitator superfamily transporter), for the GEA on the precipitation-linked variable there were 30 non-reference HOG over the 94 candidate HOG, and for the GEA on temperature associated

variables, non-reference genes represented 77 of the 217 candidates. This shows well the complementarity of reference-based SNP analysis and pangenome-based analysis. Many genes are significantly associated with the phenotypes but not annotated (such as the two unannotated candidate HOG of the pathogenic GWAS), representing a genomic resource still to be explored.

Nevertheless, these PAV-GWAS analyses have some weaknesses. Evaluating the association of phenotypes with genes, instead of SNPs may be more prone to errors. Indeed, in the SNP analysis we used the local score approach that allowed to amplify the signal on genomic regions with multiple significant SNPs in linkage disequilibrium, which can allow to reduce the number of regions falsely associated with the phenotypes. With the PAV data it is not currently possible to apply such a correction and it can be a problem, especially for the GEA in which a trend of false positive inflation was observed. Since this pangenome was mostly built with short reads data, there can be mistakes in the PAV patterns. Indeed, the absence of genes in some accessions can be due to technical issues (the accession has the genes, but the reads could not be assembled and annotated well enough to identify it), rather than to a biological reality. At the present time, our pangenome is probably more suitable for large scope analyses (enrichments for instance) than for precise PAV analysis, but this trial still gave some results that could be crossed with other analyses.

III) Concluding remarks on the pangenome

Due to the nature of the available data, we built a gene-centered pangenome, based mainly on short reads assemblies. The technical choices made here imply some limits to the exploitation of the pangenome, ordered the same way as concerns detailed by Golicz et al. **(Golicz et al., 2016)**:

- The assemblies made on short reads are not as optimal as the ones made on long reads, being way more fragmented. Thus, this approach does not allow to study large structural variations like inversions, rearrangements or indels, neither to link the additional genomic fragments not present in the reference to a location in the genome (a chromosome). We went through cleaning rounds on the assemblies, but they may not have been enough to discard the non-plant contigs from all the accessions. These contigs can bias our pangenome content, which is why we chose to focus only on annotated genes.

- Our pangenome relies on gene and can therefore be greatly influenced by the quality of the annotation. We decided to annotate our genomes based on a protein database containing a diversity of other plant proteomes already annotated, as a way of making sure we do not take into account to many genes coming from contaminations. On one side, this is a stringent approach, that can prevent us from detecting interesting genes that may not be present in the database (recent HGT, bryophyte or liverworts specific genes that have not been annotated yet). But at the same time, it still allows the potential contaminants predicted in the proteome database, to be predicted in our pangenome. I have observed first-hand the importance of the annotation process by comparing the reference annotation of the *M. polymorpha ssp ruderalis* genome with the new annotation I made of it: there was a lot of non-corresponding genes in both annotations, even though they were made on the same genome. Indeed, we used launched a large-scale annotation of our 135 genomes from diverse accessions with the same BRAKER2 pipeline **(Brůna et al., 2021)**, based on Viridiplantae gene models, whereas the reference genome was annotated through multiple steps. First Bowman and collaborators annotated the first version of the reference genome based on transcript assemblies and plant proteins alignment, using different homology-based predictors (BRAKER1 **(Hoff et al., 2016)** and FGENESH+ **(Salamov & Solovyev, 2000)**), and then refined their genes models **(Bowman et al., 2017b)**. The annotation of the current version of the reference genome is based on this first annotation (v3.1), complemented by *de novo* gene prediction obtained with BRAKER2 based on RNA-Seq librairies and the v3.1 gene models **(Montgomery et al., 2020)**. Our (re-)annotation is therefore less refined than the one from the reference genome but homogenising the annotation of all long and short-reads genomes with the same pipeline at least prevented us from falsely estimating presence absence variation that could only be due to differences in annotations.

- The detection of orthologs may be subject to errors, especially when studying heavily duplicated plant genomes. The advantage with Marchantia was that its genomes did not go through whole genome duplication rounds and is therefore lowly duplicated (only sparse tandem duplications). Indeed, when calculating the median number of genes in accessions, for each HOG of the *M. polymorpha* ssp. *ruderalis* pangenome, I could observe that most HOG (96%) only had one gene per accession, and only some HOG (1212, 4% of total HOG) contain multiple genes, going up to 27 genes at most. Many of these large size orthogroup seemed to

come from viral origin: transposon orthogroup with 15 genes, reverse transcriptase HOG with 14 genes, retrovirus zinc finger with 14 genes. Other orthogroups contained plant genes like metallothionein (10 genes), protein kinase like (8 genes) or peroxidases (7 genes). The low resolution of orthogroups is not a limiting factor for this pangenome, except maybe for duplicated gene families, like the MTPSL, peroxidases, or cytochromes p450.

- The sampling of accessions displays some biases: over sampling in specific areas (the UK, the Pyrenees), absence of sampling in the southern hemisphere. Nevertheless, it seems to cover most of the diversity of *Marchantia polymorpha*. If the work on this model species continues to develop in the following years, a better and larger sampling could be envisioned for the pangenome, the advantage with our building strategy being that new accessions can be added easily: once assembled and annotated, it only necessitates an Orthofinder run to obtain the orthogroups. Pangenomes built with the map-to-pan approach (i.e. the first approach we implemented), are not as flexible.

This pangenome probably does not offer the finest and most exhaustive vision of the genetic diversity existing in *Marchantia polymorpha*, but it still represents the first pangenome ever build in the understudied clade of bryophytes. The short reads data coming from the 135 accessions allowed comparing the gene content in the three different subspecies of the *M. polymorpha* complex. Expanding the sampling in the *polymorpha* and *montivagans* subspecies could allow to get a better insight at the genetic bases of adaptations that distinguish these three clades. The pangenome of the 100 accessions from the *ruderalis* subspecies almost captures the full diversity of *M. polymorpha* ssp. *ruderalis* and is a resource to explore the main gene families characteristic of *M. polymorpha* pangenomic compartments, and to compare their functions with the ones found in other pangenomes. In general, the core genome contains genes involved in essential processes and in Marchantia, these processes are for instance linked with DNA, RNA and protein processing (elongation factor Tu, chaperones), energy metabolism (ATPase) or plant specific mechanisms like chlorophyll degradation (pheophorbide a oxygenase). On the other side, the accessory genome is usually linked to stress response and adaptation to specific conditions. Among *M. polymorpha* accessory genes are transporters (aquaporins, phosphate transporters), that could enhance the ability of the plant to import nutrients and water, secondary metabolism enzymes (chalcone synthases or polyphenol oxidases), and a multitude of stress response proteins (NBS-LRR, pathogenesis-related proteins,

fungal fruit body lectin). The composition of the core and accessory genome of *M. polymorpha* is quite similar to what has been found in other plant species, like soybean (core genes linked to growth and cellular organisation and accessory genes enriched in biotic and abiotic stress response genes (Y. Liu et al., 2020)) or *Brachypodium dystachion* (core genes linked to essential cellular processes and accessory genes linked to defence, gene regulation (Gordon et al., 2017)). These general biological functions behind the pangenome compartments has also been observed in other organisms, like fungi or bacteria. Indeed, the study of the human pathogen *Aspergillus fumigatus* revealed that its core genes are linked to transcription, cellular protein modification and accessory genes are linked to metabolites biosynthesis (Lofgren et al., 2022)), while the comparative study of 12 bacterial pangenome showed that core genes were generally linked to metabolic and ribosomal functions, while accessory genes were associated with trafficking, secretion and defence (Hyun et al., 2022). Our Marchantia pangenome is therefore consistent with what has been found in other pangenomes, but also highlighted some Marchantia specificities. Gene families important in *M. polymorpha* adaptation have been found in its accessory genome, such as the isoprenoid (or terpenes) synthase family or the class III peroxidases. The study of the accessory genome also enabled to highlight a horizontally transferred gene that may have played a role in plant adaptation to land life: the fungal fruit body lectin.

Discussion and perspectives

The objective of this PhD was to start deciphering the adaptation mechanisms in bryophytes, by exploring the first genomic diversity dataset in a bryophyte. To do so, we used sequencing data from 135 *M. polymorpha* accessions belonging to three distinct subspecies and developed diverse resources based on this data. First of all, we developed a dense SNP dataset based on the reference Tak-1 genome v6.1, covering polymorphisms existing between the accessions with 1 SNP every 18 bp for the dataset made with all the accessions, and 1 SNP every 40 bp for the dataset built on the subspecies with the most accessions (105): *M. polymorpha* ssp. *ruderalis*. To go a bit further we also developed a gene-based pangenome by assembling and annotating long and short-reads genomes. This pangenome allows comparing the gene content of the three *M. polymorpha* subspecies with the one of their sister species *M. paleacea*. The pangenome version made only on *M. polymorpha* ssp. *ruderalis* seems to cover most of the gene diversity existing in the subspecies and gives an overview of the presence absence variation among individuals and across the different gene families. Both resources are available on the MarpolBase website (https://marchantia.info) for the Marchantia/bryophyte and plant communities to explore (Figure 76). This resource will be useful for researchers that would want to carry out comparative genomic analysis, genome-wide association studies on *Marchantia polymorpha*, as well as the ones that would want to explore a bit more the genetic diversity of their favourite candidate genes.

Figure 76: Screenshot of the SNP browser available on the MarpolBase website (https://marchantia.info/pangenome/).

With these resources built, we began to explore the diversity existing in *M. polymorpha* to get a better understanding of the genomic mechanisms behind this plant's adaptation.

I) *Marchantia polymorpha* main adaptive candidate genes

Three types of analyses have been carried out to explore the genetic bases of *M. polymorpha* adaptation. The first one was a genome-wide scan of the selection signatures left on genes, that allowed detecting genes under strong purifying selection, selective sweeps or strong balancing selection. The last type of selection maintains genetic variation within the population and plays a crucial role in organism's adaptation (D. Charlesworth, 2006; Hedrick, 2007). This is

exemplified in a study on *A. thaliana* genes under long term balancing selection, in which most candidates genes are linked with responses to biotic and abiotic stresses (Wu et al., 2017). The second type of analysis were genome wide association studies, that allow pinpointing genes correlated with a given phenotype (here the "phenotypes" being the resistance of *M. polymorpha* ssp. *ruderalis* to a fungal pathogen, or the climatic conditions at the accession sampling site). Finally, the third type of analysis was an enrichment on the functional domains present in the bryophyte's accessory genome, since this pangenomic compartment often contains genes linked to adaptation to specific conditions (Gordon et al., 2017; Y. Liu et al., 2020; Lofgren et al., 2022). These three analyses pointed to genes and gene families potentially involved in *M. polymorpha* adaptation to various conditions, with some gene families being commonly found in multiple analyses. The main gene families in this case are represented on the Figure 77 and some of them are discussed below.

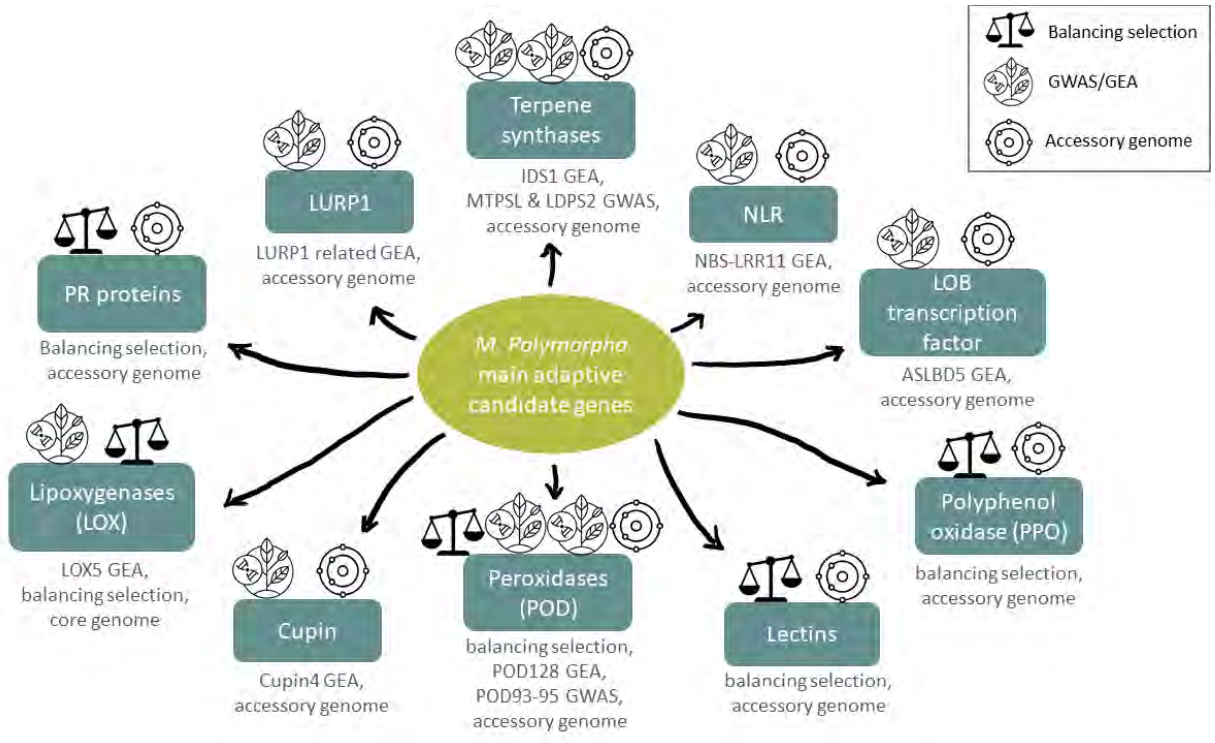


Figure 77: Main gene families appearing in multiple analyses (selection signature / genome wide association analyses / pangenome) . For GWAS and GEA candidates, the name of the gene is specified. Icons are from the Noun Project, balance icon NounProject/TkBt (CC BY 3.0), plant gene icon NounProject/Cindy Van Heerden, atom icon from NounProject/Putu Dicky Adi Pranatha.

a) The terpene synthases

The first recurrent candidates were the terpene synthases, involved in different steps of the terpene pathway. This type of protein was found in the GEA, with MpIDS1, a farnesyl

diphosphate synthase associated with temperature and precipitations variables. It was also found involved in the response of *M. polymorpha* to the fungal pathogen *C. nymphaeae*, with the GWAS candidates MTPSL and MpLDPS2. Finally, the isoprenoid synthase domain, corresponding to terpene synthase proteins was one of the domains enriched in the accessory genome of *M. polymorpha*. The sesquiterpene producing family of trichodiene synthases was also enriched in the *M. paleacea* specific genes from the Marchantia pangenome. With the large body of evidence linking this gene family to stress response in *M. polymorpha* and potentially Marchantia in general, I have investigated the terpene synthase landscape in this liverwort in order to better understand the organisation and origin of terpene synthase genes in *M. polymorpha*. It appeared that, among the two horizontally transferred MTPSL families, one was bryophyte specific (transferred from fungi) but the other one could have been transferred to the ancestor of land plants from bacteria. Most terpene synthases found here seem to be linked with *M. polymorpha's* oil bodies. Indeed, MpIDS1 is specifically expressed in oil bodies (Wang et al., 2023) and the MTPSL family found by GWAS was linked with sesquiterpene production in the oil bodies (Takizawa et al., 2021). This raises the question of the role of oil bodies in *M. polymorpha* resistance to stress. So far, this specific cell type has only been clearly linked to defence against herbivores (Romani et al., 2020) but the results found during this PhD indicate a potential role of oil bodies in the response to other biotic stresses (*i.e.* fungal pathogens) and abiotic stresses. Hypothesis concerning the role of oil bodies in these types of stresses have been formulated but never clearly proven (Romani et al., 2022), it would therefore be interesting to continue investigating oil bodies function in stress response in *M. polymorpha*.

b) The peroxidases

A second gene family well represented in our analyses were the class III peroxidases, *i.e.* the plant peroxidases. The plant peroxidase domain was found enriched in genes under balancing selection and in *M. polymorpha* ssp. *ruderalis* accessory genes. A cluster of three peroxidases (POD93, POD94 and POD95) was found associated to the response to *C. nymphaeae* and another peroxidase (POD128) is potentially associated to temperature variation. The POD93, POD 94, POD95 and POD128 candidates are all located in peroxidases clusters. The orthogroups to which these candidates belong on the Redoxibase database (Savelli et al., 2019) all seem to be originating from *M. polymorpha* or Marchantia-specific duplications. These duplications lead

to a class III peroxidase family counting 190 genes in *M. polymorpha* (according to Redoxibase), which is more diversified than plant peroxidase family in some angiosperms (75 class III peroxidases genes in *A. thaliana*, for instance). Gene duplication enables diversification of gene families, since duplicated genes often undergo relaxed selection that encourages the spread of new mutations in the population. Given that duplications are way less common in liverworts than in other land plants, the fact that class III peroxidases are highly duplicated in Marchantia suggests this gene family has a crucial adaptive role in this plant.

c) The NLRs

Another gene family highlighted by our analyses was the well-known NBS-LRR intracellular receptors that are a major class of plant disease resistance genes. NLR domains appeared in the accessory genome enrichment, some of them had missense or loss-of-function variants positively selected or under balancing selection, and the MpNBS-LRR11 was found associated to precipitations by GEA. This NLR's association with precipitation variation is either due to an indirect effect of precipitations on pathogen pressure, or could show that bryophyte NLR are involved in biotic but also abiotic stress response, as it was shown for multiple angiosperm NLRs (K. Chia & Carella, 2023). NLR were also enriched in gene families specific to *M. paleacea*, showing that *M. polymorpha* and *M. paleacea* evolved their own sets of NLR genes (approximately 24 of the 34 NLR predicted in *M. paleaceae* do not cluster in *M. polymorpha* NLR clades). The MpNBS-LRR11 candidate that appeared in the GEA and in the variant analysis is a coil-coil NLR, which is a common type of NLR present in all land plants that triggers a calcium ion influx through the plasma membrane and lead to hypersensitive-response (HR) cell death. I carried out a preliminary analysis on all proteomes from the *M. polymorpha* pangenome which showed that only a few *M. polymorpha* NLR were well-known plant NLR (TIR-NLR, coil-coil NLR or RPW8-NLR) and that these NLRs seemed highly conserved between accessions. On the opposite, the most common Marchantia NLRs are αβ-hydrolase NLRs (HNL), an atypical NLR domain present in liverworts and in mosses (Andolfo et al., 2019; K.-S. Chia et al., 2022). *M. polymorpha's* HNL seem highly variable, with presence absence polymorphism, variable gene size and variable domains between accessions. With respect to variability, Marchantia's HNL seem to undergo similar diversifying mechanism as common NLRs (TNL, CNL, RNL) in angiosperms. The 39 NLR identified in the *M. polymorpha* ssp. *ruderalis* reference genome show that the total number of NLR encoded in the genome is quite different between

Marchantia and most angiosperm (NLR numbers in angiosperms are quite heterogenous, ranging from 5 to more than 2000, with only 30 over 305 angiosperm genomes evaluated counting less than 50 NLRs (Y. Liu et al., 2021)). This can be explained partly by the absence of whole genome duplication and low genome redundancy of *M. polymorpha*. But other adaptive gene families, such as the peroxidases, still underwent significant diversification in the liverwort. A hypothesis could be that NLR are an important part of *M. polymorpha's* response to biotic stresses, but that other gene families such as the peroxidases or the terpene synthases are the main ones recruited during the stress response. Angiosperms display an important diversity of secondary metabolites and liverworts developed NLR families, but there could be two distinct strategies for biotic stress response in these two lineages of land plants: one mainly centred around NLRs, in angiosperms, and one mainly centred around a rich panel of secondary metabolites in liverworts (Asakawa & Ludwiczuk, 2018). This difference in NLR family size could also be due to different ecological strategies between most angiosperms and liverworts, since angiosperms that adopted alternative lifestyles (reversion to water, carnivorous, parasitic) have less NLR (Y. Liu et al., 2021).

d) The pathogenesis related (PR) proteins

The very general category of pathogenesis related (PR) proteins refers to components of the plant innate immune system, that often have antimicrobial activities, and that can be classified into 17 categories. PR proteins have also been shown to increase plant resistance not only to biotic but also to abiotic stresses (Ali et al., 2018). Different types of these proteins were found in our analyses. PR1 proteins, that are small cysteine rich secreted proteins that have broad antimicrobial activity and amplify the defence signal (Breen et al., 2017), have been found enriched in genes under balancing selection and in accessory genes. Some PR5 proteins, that are thaumatin-like proteins, are under shared balancing selection in *M. polymorpha*, *A. thaliana* and *M. truncatula*. Chitinases like the GH17 enriched in the accessory genome can be PR proteins (from the PR2, PR3, PR8 or PR11 family) (Ali et al., 2018). Finally, peroxidases, that were mentioned previously, are also classified as PR9 proteins and cupins as PR15/PR16 proteins.

Various types of pathogenesis proteins therefore seem to have an important role in the adaptation of *M. polymorpha* to different biotic and potentially abiotic conditions. This is

consistent with the induction of PR proteins found by Carella and collaborators (Carella et al., 2019) when studying the infection of *M. polymorpha* by the oomycete *Phytophthora palmivora*.

e) The lectins

Another large family of plant proteins was found linked to Marchantia's adaptation: the lectins. These are carbohydrate binding proteins mainly involved in plant immunity and symbiosis, but that can also be involved in plant development and stress signalling. Plant lectins can be distinguished in 12 different families based on their lectin domains (De Coninck & Van Damme, 2021). In our analyses four lectin families appeared: the bulb type lectins (or *Galanthus nivalis* lectin family) that appeared enriched in the accessory genome and among genes under balancing selection, the ricin B lectin family and the concavalin-A-like lectins (legume lectin family), that were also enriched among genes under balancing selection. Finally, the fungal fruit body lectins (or *Agaricus bisporus* agglutinin family) were found enriched in the accessory genome. Except for the fungal fruit body lectins, that originated in fungi and were then passed to land plants, all lectins described here probably originated in the last universal common ancestor of Bacteria and Archaea, and are therefore present in some green algae (Van Holle & Van Damme, 2019).

f) The lipoxygenases

Lipoxygenases (LOX) have also been found in diverse analyses: they are enriched among the genes under balancing selection, the "plant lipoxygenase" category is enriched in *M. polymorpha* core genome, and a lipoxygenase was found correlated to precipitations in the GEA. Lipoxygenases catalyse the oxidation of polyunsaturated fatty acids to oxylipins, that lead to a variety of compounds, such as antimicrobial compounds or jasmonates (Feussner & Wasternack, 2002). In land plants two hormones interact to mediate the plant response to biotic stresses: salicylic acid (SA) that mediates the response to (hemi)biotrophic pathogens, and jasmonates (JA-Ile in tracheophytes or dn-ODPA in bryophytes) that mediates the response to herbivores and necrotrophic pathogens (Monte, 2023). These two phytohormones are also linked to plant development, fertility and response to abiotic stresses. Thus, LOX play a role in *M. polymorpha* defences and stress response through hormone signalling pathways, as they do in other land plants.

All these candidates are already well known in plants, which shows that Marchantia shares an important part of its stress response mechanisms with other land plants. Nevertheless, even though these gene families are common in plants, *M. polymorpha* displays some peculiarities in the genes it uses. The terpenes synthases and some lectins are of a microbial origin and are not the classical plant families, the peroxidases are mostly Marchantia-specific because they have been highly duplicated in this lineage, and the NLR are mostly alpha beta hydrolase NLR (HNL) that do not even exist in vascular plants.

Most of these gene family expanded in *M. polymorpha* seem to be tandemly arrayed genes (TAGs), that are not uncommon in Marchantia: there is 5.9% of TAG, which is among small values compared to angiosperms (range from 4.6% to 26%), but higher than the moss *Physcomitrium patens* (1% TAG) (Bowman et al., 2017b). Given the low level of duplication in Marchantia's genome, genes in tandem arrays are indicators of adaptive gene families for which selective forces promoted diversification.

The candidates presented here are, of course, not an exhaustive list of the genes involved in the *M. polymorpha's* response to stresses. For instance, genes related to the phenylpropanoid biosynthesis pathway that play a role in Marchantia's defence against an oomycete pathogen (Carella et al., 2019), were not among the main candidates of our analyses. The genome-wide analyses give large lists of candidate genes that need time to be investigated, and there can be a bias towards well known gene families because they will attract the attention of the human curator, but also more generally because they are better annotated than many other genes. In the enrichment analyses, results can also be biased towards larger gene families. Nevertheless, these analyses granted us with a list of potential candidate genes that could be further investigated through functional validations.

II) *Marchantia polymorpha* population genomics characteristics

Our genomic diversity dataset did not only allow finding candidate gene families that could be involved in *M. polymorpha's* response to various conditions. Its study also gave rise to questions concerning multiple aspects of the genomic structure and evolution mechanisms in this bryophyte.

For instance, the genome-wide association study on gene PAV led me to consider the extent and functional importance of presence absence variation and more generally structural variation in the genome of Marchantia, compared to other plant species. Indeed, the importance of presence absence variation in the genetic architecture of traits has been demonstrated in angiosperms, but these species often have more complex genomes than *M. polymorpha*, like soybean that underwent multiple WGD (Y. Liu et al., 2020) or the allotetraploid crop *Brassica napus* (J.-M. Song et al., 2020). PAV may be more predominant in this type of species, than in *M. polymorpha* that did not undergo any whole genome duplication and has a low TE and repeats content (27% of the genome in Marchantia, as opposed to 56% in *Physcomitrella patens'* genome (Lang et al., 2018; Montgomery et al., 2020)). These phenomena promoting large structural variation (Saxena et al., 2014) are scarce in *M. polymorpha*, maybe due to the presence of strongly dimorphic sex chromosomes which complicate meiosis in tetraploids (Bowman et al., 2017b). PAV may therefore have less impact on the phenotypic traits of this bryophyte than on the traits of some angiosperms.

This comparison between the highly complex and duplicated genomes of angiosperms and the low redundancy in *M. polymorpha's* genome raises the question of the relevance of genome simplicity and reduction in evolution. Contrary to the common belief, evolution on a broad phylogenetic scale does not always tend toward increased complexity (be it at the genomic or organismal level), but might consist of two phases: an innovation phase during which genome complexity of organisms increases abruptly through macromutations, and then a longer reductive phase, in which genetic material is lost and reshuffled via adaptive or neutral processes (Figure 78) (Wolf & Koonin, 2013). Since the speciation burst of the liverwort lineages happened at least 100 MY before the one of angiosperms and gymnosperms (Figure 6), it could be envisioned that Marchantia and other liverworts might already be well advanced in the genetic reductive phase of evolution, compared to younger lineages like spermaphytes (seed plants). In the biphasic model of punctuated evolution of genomes presented by Wolf and Koonin (Wolf & Koonin, 2013), the genetic modifications happen at a slower rate during the simplification phase than during the complexification phase, which correlates with low silent site substitution rates (compared to angiosperms (Linde et al., 2021)), the low levels of gene duplication, TE activity, chromosomal rearrangement and probable morphological evolution observed in liverworts (Linde et al., 2023). Concerning other bryophyte lineages, hornworts

might be in the same reductive phase since they also display low genomic redundancy (J. Zhang et al., 2020), whereas mosses seemed to have gained a bit more gene families than they have lost, probably because they underwent multiple rounds of whole genome duplication (Lang et al., 2018; J. Zhang et al., 2020).

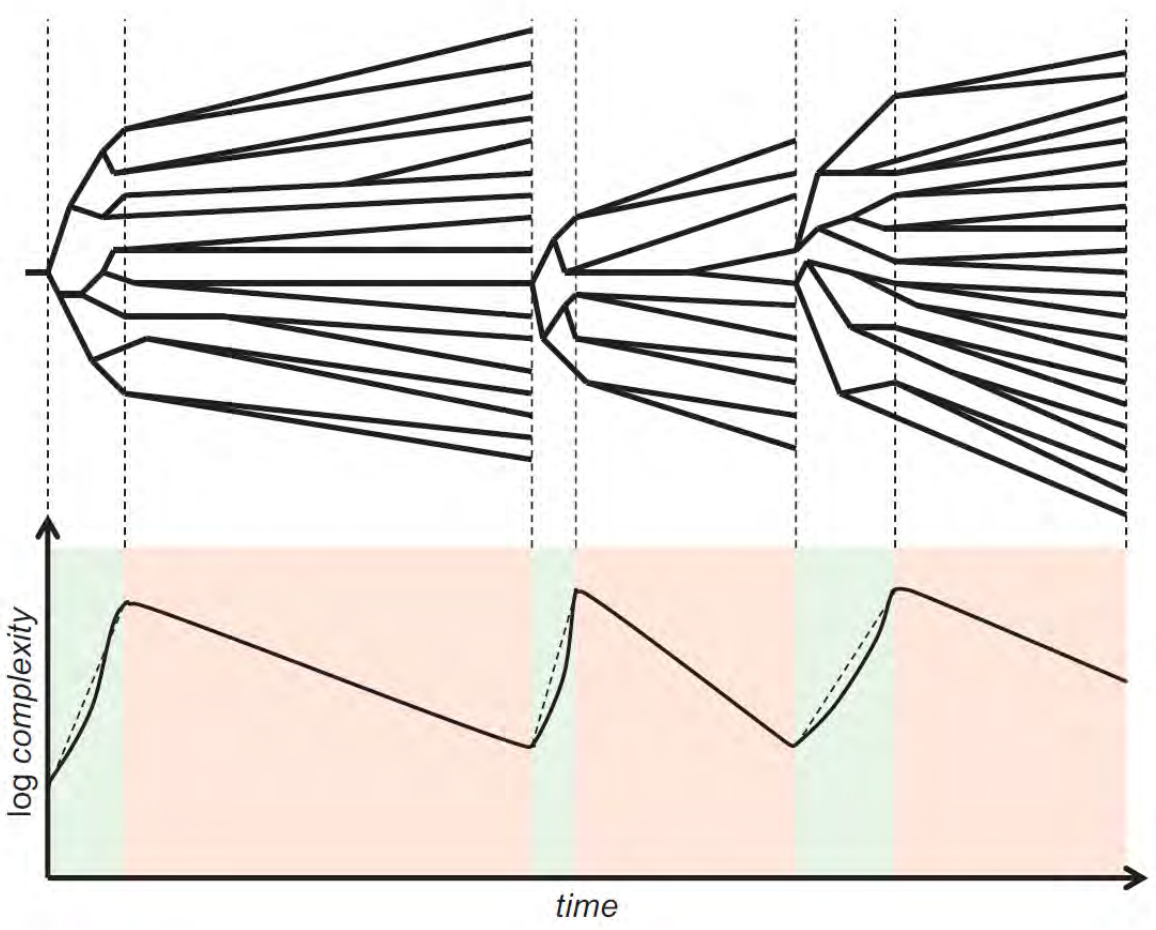


Figure 78: Biphasic model of punctuated evolution as presented by Wolf and Koonin. On the upper panel is represented the periods of cladogenesis followed by long phases of quasi stasis in the history of lineages. On the lower panel is the complexity profile corresponding to this cladogram. The complexity (that can be expressed as the number of sites or genes that are subjected to selection) increases during brief periods of time (in green, the complexification phase) and then steadily decreases during extended periods of time (red, reduction phase). Figure retrieved from Wolf and Koonin 2013 ***(Wolf & Koonin, 2013)****.*

In the framework of adaptation (because changes in genome complexity can also happen via non-adaptive processes like genetic drift), the succession of genomic complexification and simplification enables selective evolutionary processes. Genomic complexity is then a base on which selection can act via adaptive gene loss to produce efficient functions governed by a reduced set of genes, following the "less is more" hypothesis (Monroe et al., 2021; Olson, 1999; O'Malley et al., 2016). Many loss-of-function mutations (LoF) have been shown to play a role in

adaptation, such as the null mutation in the human CCR5 gene that confers resistance to HIV (Olson, 1999) or the frequent loss of function alleles found in genes linked to the adaptive trait of seed dormancy in some *A. thaliana* populations (Xiang et al., 2016). In our *M. polymorpha* ssp. *ruderalis* population, the importance of LoF mutations was confirmed by the start lost, stop lost, stop gained and other LoF mutations that were almost fixed in all accessions, and often located in adaptive genes.

In the case of Marchantia, the LoF mutations (and other mutations) are directly exposed to selection, since the plant's life cycle is dominated by the haploid gametophyte, which contrasts with most studied angiosperms which show a dominant sporophytic (diploid or even polyploid) phase in their life-cycle. The masking hypothesis predicts that purifying selection, as well as positive selection, would be more efficient in haploid species since mutations are not masked by a second allele (Otto & Goldstein, 1992). In previous studies no clear difference in selection pressure (measured by the ratio of non-synonymous to synonymous substitutions, d_N/d_S) was detected between angiosperms and haploid bryophytes (Linde et al., 2021). However, these estimations of selection pressures can be biased (low number of genes considered, alignement issues) and the study of intraspecific genomic diversity in these predominantly haploid organisms could improve our understanding of the evolutionary dynamics of deleterious and adaptive mutations in bryophyte species

The haploid genome of Marchantia also has a consequence on the nature of balancing selection at work on its genes. In a lot of studied organisms, balancing selection results from a heterozygote advantage, such as the one conferred by the sickle cell allele at the β-hemoglobin locus in humans, that grants heterozygotes with resistance to malaria. Marchantia is haploid and, even during its short-lived diploid phase, the paternal genome was shown to be epigenetically repressed (Montgomery & Berger, 2023). Thus, there can theoretically be no selection driven by heterozygote advantage in *M. polymorpha.* Therefore, other mechanisms should explain patterns of balancing selection identified in several *M. polymorpha* genes and genomic regions. These patterns could be due to frequency-dependent selection, in which rare alleles confer an advantage, or temporally or spatially heterogeneous selection that would occur across the *M. polymorpha* collection (D. Charlesworth, 2006). Temporal heterogeneity could be caused for instance by fluctuating pathogen populations, or by spatial heterogeneity due to the sessile nature of plants and by the importance of microhabitats.

Another question of interest concerning *M. polymorpha's* genome is the importance of sexual and asexual reproduction in its reproductive strategy. Indeed *M. polymorpha* is a dioecious plant that can reproduce via sexual mating, or vegetatively, via its gemmae. No evidence of direct clonal descent and signatures of frequent outcrossing and long-distance gene flow have been found with a local collection from Southern Ontario (Sandler et al., 2023), which is consistent with the long-distance gene flow and low linkage disequilibrium we observed in our worldwide collection. Only our highly local populations seemed clonal (the Tou or Bul populations for instance, Figure 18). Considering these patterns of low linkage disequilibrium and long-distance gene flow, it seems that sexual reproduction is actually quite common and enabled by a gamete dispersal on longer distances that what was initially though, which could be enabled by animal-promoted gamete dispersal, as it has been observed in mosses (Shortlidge et al., 2021). Transport of individuals over long distances through horticultural trade (Marble et al., 2017) can also enable reproduction in initially distant plants.

Another interesting question linked to *M. polymorpha's* reproduction is the amount of genetic diversity introduced during clonal reproduction, via somatic mutations. This could be studied, for instance, by comparing multiple genomes from the TAK1 reference accession cultivated independently and reproduced clonally in different labs around the world. The accession TAK1-HK from our collection was sequenced for this purpose, but unfortunately the sequencing was of poor quality and did not allow comparison with the reference genome. A sequencing of accessions in the centre and in the periphery of clonal local populations (unisex patches of Marchantia) could also allow to study the apparition of these somatic mutations in nature, similarly to the sequence comparison of various leaves from a 234 years old oak tree (Schmid-Siegert et al., 2017).

Finally, the contribution of adaptive introgressions between the three *M. polymorpha* subspecies is still to be examined using the genomic resources generated during this project. In previous works, crossing experiments were carried out leading to viable spores (only for the cross between a female from the *montivagans* subspecies and a male from the *polymorpha* subspecies (Burgeff, 1943)), showing that crosses between some individuals from the three subspecies might still be possible, although the three subspecies have diverged around 5 MYA. In our collection, the accession Pra-B could be a hybrid between the *ruderalis* and *polymorpha* subspecies, given his phylogenetic position. Linde and collaborators (Linde et al., 2020) studied

gene flow between *M. polymorpha's* subspecies on a smaller sample of accessions and found signs of frequent hybridization and introgressions. To further explore introgression patterns in *M. polymorpha* using our genomic resources, a new method is under development by collaborators at the SETE institute (Hervé Philippe & Simon Aziz, Moulis, France) to quantify genomic introgression events within *M. polymorpha ssp ruderalis* and among *M. polymorpha* subspecies. This method uses an empirical approach calculating window-based SNP p-distances within and between subspecies. Once the method and the list of introgressions detected in the subspecies will be finalised, we will have a better understanding of the introgression landscape in *M. polymorpha*, of its importance and of the gene functions underlying the adaptive gene exchanges that may have occurred between *M. polymorpha* subspecies.

III) Perspectives on interspecific comparisons

Studying *M. polymorpha*, we also ambitioned to get a better understanding of the mechanisms of adaptation existing in a diversity of land plants, by comparing our findings with angiosperm data. By studying a diversity of species, and not only the classical model plants or main crops, it will be possible to expand our knowledge of the adaptive mechanisms and the evolutionary genomics of land plants. Comparisons across a range of species enable to project the results into an evolutionary perspective, and to either generalise observations or define their specificity within a single species/taxonomic space. In our case, it allowed to identify commonalities in the genetic strategy for adaptation to the environment in land plants, but also lineage-specific innovations deployed by *M. polymorpha*. This was exemplified by the candidate horizontal gene transfers found in the accessory genome of *M. polymorpha* and with the GWAS analysis. This fungal fruit body lectin and microbial terpene synthase-like genes were already known in *M. polymorpha*, but our phylogenetic analysis revealed that they are also present in vascular plants genomes. Previous studies mentioning these two types of HGT though they were clade specific because they lacked the lycophyte and monilophyte genomes we have used in our study (Jia et al., 2016; Kumar et al., 2016; Peumans et al., 2007; Van Holle & Van Damme, 2019). These new results change the whole nature of these horizontal gene transfer, from bryophyte-specific HGT to transfers that may have helped the land plant MRCA to adapt to its new terrestrial environment (Ma et al., 2022).

Apart from the projection of some *M. polymorpha* candidate genes in a macro-evolutionary context, we also tried to go further into the cross-referencing of intra-specific and inter-specific data. This comparison across evolutionary scales was performed by crossing the intraspecific selection signatures left on genes in three different species: *M. polymorpha*, *M. truncatula* and *A. thaliana*. This comparison came with challenges, in particular concerning the multiple whole genome duplications that occurred during the evolutionary history of angiosperms. Orthogroups with similar selection signatures in the three species often contained many genes from the two angiosperms species, and a single, or few, pro-orthologs from Marchantia. It was thus difficult to determine if any common selection signature occured on genes with equivalent biological functions or not, considering potential sub- and neo-functionalizations.

I also tried another interspecific comparison, based on presence-absence variation in pangenomic studies from multiple species (cucumber (H. Li et al., 2022), *Medicago truncatula* (P. Zhou et al., 2017), pigeon pea (J. Zhao et al., 2020), rice (Q. Zhao et al., 2018), sunflower (Hübner et al., 2019), and *M. polymorpha*). I performed GO term enrichment on accessory and core compartments from all these pangenomes, in order to determine if there were common enriched terms for accessory and core genes in these diverse plants. By comparing the different enrichment no term common to all the core genomes or all the accessory genomes was detected (example for the enrichment of accessory genes in Appendix G). The size and nature of the sampling, the nature of the sequencing data, the assembly and annotation strategies, and the construction strategy for the pangenome were all parameters that made this data very heterogeneous and therefore hard to precisely compare. Nevertheless, we were still able to cross the general molecular functions found in our pangenomic compartments with the ones enriched in other pangenomes and to assert that the classical characteristics of core (essential) and accessory (stress responsive) genome were also present in *M. polymorpha's* pangenome.

Even though genome-wide interspecific comparisons are not always easily tractable, a long term objective that motivated the construction of this dataset was to explore the interaction between symbiotic and parasitic relationship on the evolution of plant genomes (Mic-Mac project: exploring interaction between mutualism and parasitism at micro and macro-evolutionary scales). Indeed, symbiotic and parasitic microorganisms are two extremes on the same continuum of plant-microorganisms interactions (Delaux & Schornack, 2021), and might dialogue with the same plant proteins (such as the RAD1 transcription factor of *M. truncatula*

which is essential for the arbuscular mycorrhizal (AM) symbiosis and provide susceptibility to *P. palmivora* (Rey et al., 2017)). It would therefore be interesting to study the interconnection between plant immunity networks and symbiotic pathways, to test whether the retention or loss of symbiosis affects the evolution of the plant immune system. The study of these mechanisms relative to the AM symbiosis would be possible by developing a population genomic diversity dataset in another bryophyte species able to form AM symbiosis (like *Lunularia cruciata* or *Marchantia paleacea*, for instance). Comparisons of selection signature on genes and gene PAV could then be carried out across the land plant phylogenetic tree, in vascular and non-vascular plants, depending on their symbiotic status. This would allow pinpointing genes with contrasting evolutionary constraint between symbiotic and non-symbiotic plant lineages.

More generally, comparative population genomics approaches across land plants will allow to test hypotheses on the evolutionary processes acting to constrain, shift or diversify gene polymorphism throughout the land plant phylogeny, in relation with the loss, maintenance and diversification of biological traits during evolution.

Carbon footprint

of the PhD project in the context of the climate crisis

Carbon footprint of the PhD project

I used some features of the GES1.5 tool from labo1.5 **(Mariette et al., 2022)** in order to evaluate the approximated carbon footprint linked to my PhD. I took into account the computing equipment I used, the two conferences I attended to, my commute to work, the energy consumption at my workplace and calculation time I used on the calculation cluster (cf SupData4.0).

For the commute to work and mission travels I used the estimation of the GES1.5 tool as it is. For other estimations, since GES1.5 is designed to estimate emissions of whole labs, I pondered the CO2 equivalent (eCO2) by the time I stayed or space I used.

For the energy consumption, I used the gas and electricity consumption of the lab in 2022 that I pondered by the approximated space I use in the lab and multiplied by 3.25 (the duration of my PhD contract).

For the eCO2 of the computing devices I used, I took the eCO2 of these objects estimated by GES1.5 and divided them by their approximated lifespan and multiplied by 3.25.

I added a very rough estimation of the carbon footprint of my use of the computing cluster of the genotoul bioinfo platform, accounting for the use of 100 000 computing hours the first year, 150 000h the second year and 100 000h the third year, leading to a eCO2 of 1 925 000 g (https://bioinfo.genotoul.fr/index.php/news/newsletter-35special-carbon-footprint-issue/).

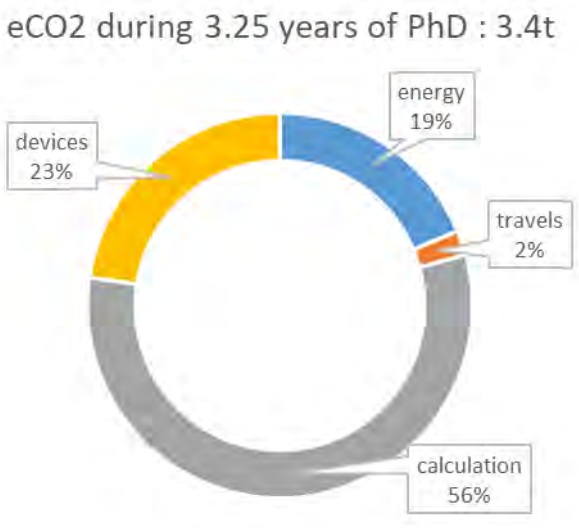


Figure 79: Proportions of eCO2 emissions for each carbon emission factor as part of my PhD project

These calculations combined, led to an estimated 3.4 tons of eCO2 emitted during my 3.25-year contract. The carbon footprint per year and per person to have a chance of limiting warming to 2°C in 2100 is of approximately 2 tons of eCO2 per year and per person (Intergovernmental Panel On Climate Change (Ipcc), 2023), and my total eCO2 budget as part of my PhD has been of 1.04 t per year, which represents already half of my personal eCO2 budget. Even though this estimation is quite rough, it still gives a good idea of the problem we are currently facing. I have made a lot of mistakes launching computing jobs that could have been avoided, and it is important to be reminded of the cost of all these invisible operations (the initiative of the Genotoul Bioinfo team to display the eCO2 consumption on the cluster is already a very nice step!). But in system where we are always compelled to produce more science, faster, it is sometimes hard to take a step back and weight the consequences of our experiments beforehand.

During the past three years I spent in the academics, some people have brought up the discussion about the role of researchers in the face of the upcoming struggles, given that scientists are the ones characterising the causes and consequences of the human-caused climate change and biodiversity collapse. I have been greatly impacted by the reflexions around the idea that maybe the scientific community should be the one setting an example, and therefore that research should be shaped to emit a minimum amount of eCO2, and that sometimes choices should be made among research topics. Another non-negligible responsibility of scientists is the transmission of these ideas and even more broadly of scientific subjects to the population, ideally to have a society where everyone would have a decent degree of understanding on the complex mechanisms our activities trigger. These are thorny questions that will remain open, but definitely need to be considered in our everyday activities given the emergency of the situation.

List of figures and tables

List of figures

Figure 1: Summary of the genetic diversity drivers. 5
Figure 2: Illustration of the mixed linear model that can be used to test the association of a SNP with a quantitative phenotype of interest..... 14
Figure 3: Illustration of the local score approach. 15
Figure 4: Representation of the concepts of open and closed pangenome, and core and accessory pangenomes. 18
Figure 5: Different pangenome construction approaches. 19
Figure 6: Phylogenetic relationships between clades of the green lineage. 25
Figure 7: Illustration of a potential evolutionary scenario for the conquest of land by streptophytes.. 26
Figure 8: illustration of some of the key features that enabled plants to conquer land. 28
Figure 9: Phylogenetic context of Marchantia polymorpha, plotted along the geological timescale from (Bowman et al., 2022). 33
Figure 10: Transverse section of the thallus showing the dorsal and ventral characteristics of M. polymorpha. Retrieved from (Shimamura, 2016). 34
Figure 11: Life cycle of M. polymorpha from (Shimamura, 2016). 35
Figure 12: Summary of the PhD project, highlighting the main question and the approaches used to address it. 38
Figure 13: Sampling location for the accessions of M. polymorpha that constitute the collection. 42
Figure 14: City marchantia, country marchantia..... 43
Figure 15: Synteny between the reference genome (TAK1) and the long reads assemblies of BoGa (left) and CA (right). 47
Figure 16: Linkage disequilibrium decay in Marchantia polymorpha, for each chromosome, and for the whole genome (black line). 50
Figure 17: Geographic distance between accessions vs genetic distance between accessions. 51
Figure 18: Phylogenetic tree of the accessions from the M. polymorpha collection, with their geographical origin. For the accessions from the ruderalis subspecies, the assignation to three different genetic groups by a FastStructure analysis is represented on the top of the tree..... 53
Figure 19: Allele frequency spectrum for different categories of SNP, based on minor allele frequency. 55
Figure 20: Detail of the different type of selective forces and their effect on the frequency spectrum in a genomic region. 56
Figure 21: Empirical distributions of Tajima's D, Fay and Wu's H and Zeng's E (top, Source: Léa Boyrie's PhD manuscript) and real distribution in M. polymorpha ssp. ruderalis based on 18 140 genes for D and 17 027 genes for H and E (bottom). 59
Figure 22: comparison of the D, H and E statistics calculated on the simulated loci by the window-based approach (x axis) and on the same loci with 50% NA, with the averaging-based DHE calculation that implements a missing data correction (y axis). 62
Figure 23: Allele frequency spectrum for different categories of SNP, based on derived allele frequency. 66
Figure 24: Barplot representing the IPR domains enriched in M. polymorpha's genes under background selection..... 68
Figure 25: Barplot representing the IPR domains enriched in M. polymorpha's genes under balancing selection..... 69
Figure 26: Distribution of the accessions along the PC1 and PC2 of temperature linked bioclimatic variables, and contribution of the variables to the construction of the PC. 78
Figure 27: Distribution of the accessions along the PC1 and PC2 of the precipitation linked bioclimatic variables, and contribution of the variables to the construction of the PC. 79

Figure 28: Genomic regions in M. polymorpha associated with the 2 first principal components of the PCA on temperature and precipitation bioclimatic variables. 82
Figure 29: Heatmap representing the SNP matrix in the region of the GEA peak surrounded by the ABC1K and the cytochrome c1. 83
Figure 30: Distribution of the accessions with the minor (yellow) and major (blue) haplotype in the ABC1K genomic region..... 84
Figure 31: Orthologs of the ABC1K. 85
Figure 32: Overview of the isoprenoid biosynthetic pathway in plants with a focus on A. thaliana's genes (FPS1 and FPS2), adapted from (Manzano et al., 2016). 87
Figure 33: : Heatmap representing the SNP matrix in the region of the GEA peak surrounded by the NLR and the LURP-1 related protein. 88
Figure 34: Heatmap representing the SNP matrix in the region of the GEA peak surrounded by the RNA II polymerase mediator and the HSP20. 89
Figure 35: Heatmap representing the SNP matrix in the region of the GEA peak surrounded by the LOB domain protein and the unannotated protein. 90
Figure 36: Heatmap representing the SNP matrix in the region of the GEA peak surrounded by the peroxidase, ribosomal protein and dynamin. 92
Figure 37: Position of the accessions relative to the PC1 and PC2 of the PCA on temperature (left) and precipitation (right) bioclimatic variables. 95
Figure 38: Experimental design used for the phenotyping of the 87 accessions of M. polymorpha inoculated with C. nymphaeae. 97
Figure 39: Variation of the phenotypes from the 87 M. polymorpha accessions: thallus area of non-inoculated plants on the upper part, and brown area of inoculated plants on the lower part.The raw measures are represented by the boxplots and the corrected means are symbolised by the diamonds. 99
Figure 40: QQplots of the p-values obtained for the 5 phenotypes evaluated in the GWAS. 101
Figure 41: Manhattan plots of the GWAS on the thallus area of the non-inoculated (upper plot) and inoculated (lower plot) plants at 6dpi. 103
Figure 42: Detail of the alleles on the SNPs located close to the peak of the trihelix transcription factor (chromosome 3). 104
Figure 43: Manhattan plot of the association with the thallus area difference between infected and non-infected plants. 106
Figure 44: Accession position relative to the two first components of the PCA on the five phenotypes from the GWAS. 107
Figure 45: Manhattan plots of the GWAS on M. polymorpha's response to fungal infection at 6dpi. 108
Figure 46: Detail of the alleles on the SNPs located close to the peak of the receptor like kinase (chromosome 2). 109
Figure 47: Detail of the alleles on the SNPs located close to the peak of the MTPSLs (chromosome 6). 110
Figure 48: Detail of the alleles on the SNPs located close to the peak of the MpLDPS2 (chromosome 8). 111
Figure 49: Landplant orthologs of the MpLDPS2 gene. 112
Figure 50:Detail of the alleles on the SNPs located close to the peak of the peroxidases (chromosome 5). 113
Figure 51: Sequence similarity network of M. polymorpha's terpene synthase genes. 114
Figure 52: Role of the terpene synthesis related enzymes found in Marchantia in the terpene biosynthetic pathway (adapted from (Habtemariam, 2019). 115
Figure 53: phylogenetic tree of the 16 MTPSL family. 117

Figure 54: phylogenetic tree of the GWAS MTPSL family. 118
Figure 55: extract of M. polymorpha's genome browser, showing the transposon signature flanking all the genes under the GWAS peak. 119
Figure 56: Details of the pipeline used for the iterative assembly of M. polymorpha ssp. ruderalis pangenome for 58 accessions. 124
Figure 57: Distribution of the genes predicted in the pangenome (reference genome initial annotation and additionnal scaffolds annotation with BRAKER2), depending on their presence in a given number of accessions. 126
Figure 58:Distribution of the additional contigs of the pangenome, depending on their GC content, and with the information of their length in bp. 127
Figure 59: Distribution of the 27-mers from the CA accession, according to their frequency of occurrence in the libraries and their GC content. 129
Figure 60: GC and length profile of contigs for two accessions, before and after the cleaning processes. 132
Figure 61: Overlap between the HOG present in at least one accession of the species. 135
Figure 62: Landscape of gene presence/absence variation in the Marchantia genus. HOGs are sorted by their occurrence, with the genes shared by all the Marchantia species in the left part (separated from the rest by a dotted line), and the very rare genes in the right part of the matrix. 137
Figure 63: Modelling of the pangenome expansion (green) and of the core genome reduction (orange) when adding accessions. 139
Figure 64: Distribution of the HOG depending on the number of accessions they contain and corresponding pangenomic compartments. 140
Figure 65: Assignation of the proteomes of each accession to the pangenomic compartments. 141
Figure 66: Selection of enriched IPR domains (FDR q-value <0.05) in the core genome of M. polymorpha ssp ruderalis. 143
Figure 67: Selection of enriched IPR domains (FDR q-value <0.05) in the accessory genome of M. polymorpha ssp ruderalis. 146
Figure 68: Phylogenetic tree of the orthologs of the fungal lectin genes present in the reference genome of M. polymorpha ssp ruderalis. 148
Figure 69: RNAseq conditions for which the core or accessory genome are significantly more up regulated (upper part) or down regulated (lower part). 150
Figure 70: Expected and observed p-values (QQplot) of the association of HOG displaying PAV in the 70 accessions in the ssp. ruderalis pangenome with the variation of response to C. nymphaeae infection. 153
Figure 71: Presence (in blue) or absence (in yellow) variation of the different candidate HOG in the phenotyped accessions. 154
Figure 72: QQplots for the two first principal components of precipitation related variables in the 91 accessions of M. polymorpha ssp. ruderalis. 156
Figure 73: matrix of the presence absence variations in 90 accessions of 50 candidate HOG correlated with precipitation variation between sampling sites. 158
Figure 74: QQplots for the two first principal components of temperature related variables in the 91 accessions of M. polymorpha ssp. ruderalis. 159
Figure 75: matrix of the presence absence variations in 90 accessions of the 91 candidate HOG correlated with temperature variation between sampling sites. 161
Figure 76: Screenshot of the SNP browser available on the MarpolBase website. 168
Figure 77: Main gene families appearing in multiple analyses (selection signature / genome wide association analyses / pangenome) 169
Figure 78: Biphasic model of punctuated evolution as presented by Wolf and Koonin. 176

Figure 79: Proportions of eCO2 emissions for each carbon emission factor as part of my PhD project 184

List of tables

Table 1: General information for the 16 accessions from the montivagans subspecies, for the 14 accessions from the polymorpha subspecies, and for the 105 accessions from the ruderalis subspecies. 43
Table 2: Detail of all the variable tested in the genome environment association study, and of the number of significant genomic regions associated 80
Table 3: Names of the phenotypic variable used in the GWAS analysis and detail of how they were obtained 96
Table 4: Details of the hits for all the genes present in the additional contigs, in the different domains of life, also including viruses. 130

.

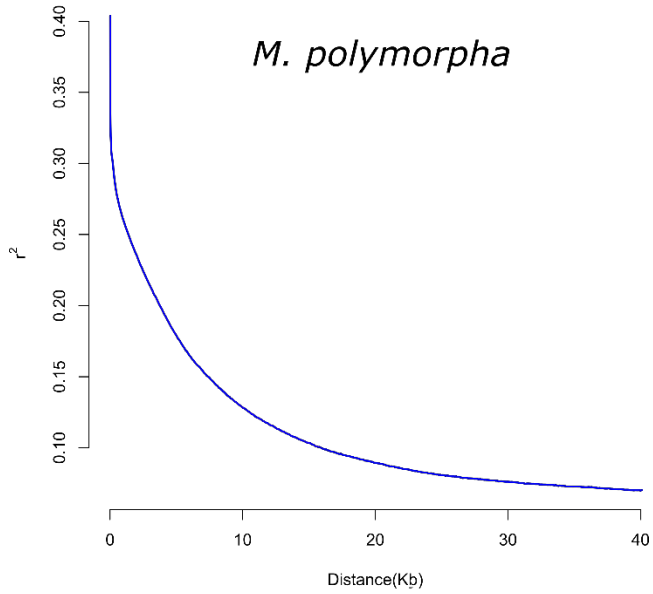
Appendix

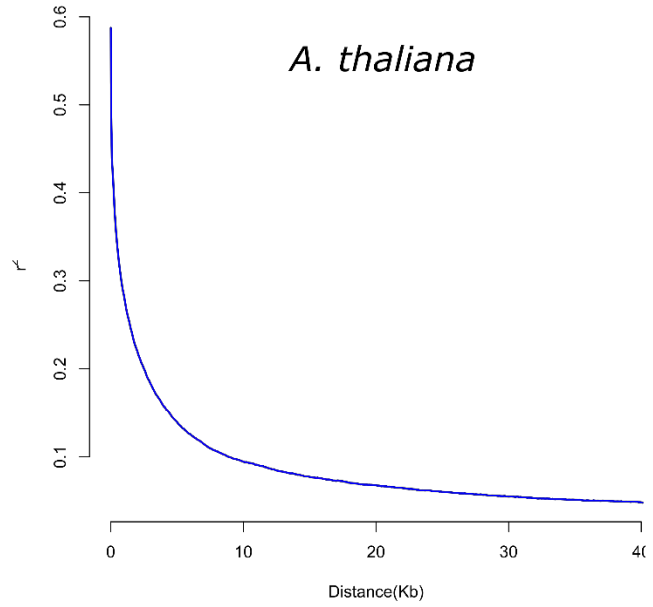
Appendix

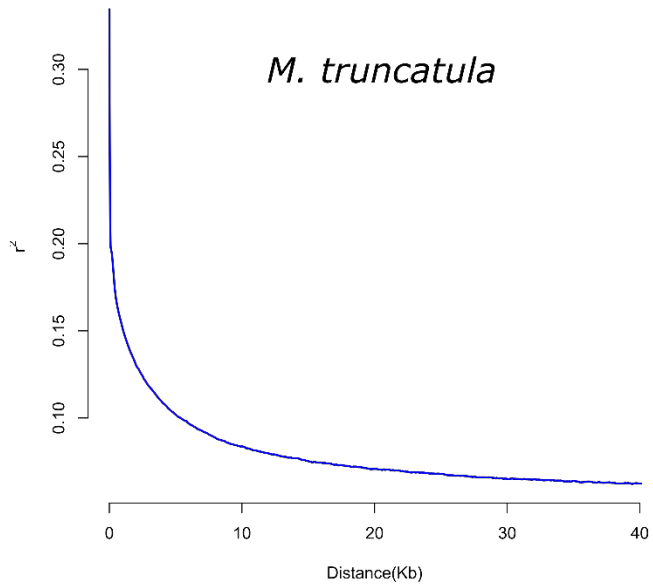
All the documents referred as “SupData” are available at the following link: https://figshare.com/s/5e00d9327d1530bfe14f. The vectorised versions of the haplotype illustration for the GWAS and GEA candidates are also available at this link, under the ./chap2/all_heatmap_haplotype/ folder.

The script that allows calculating the D, H and E statistic estimators is available at the following link: https://figshare.com/s/715a36bc7585d46d0279 (folder DHE_calculation_scripts).

Figures referred to as “Appendix” are hereunder.

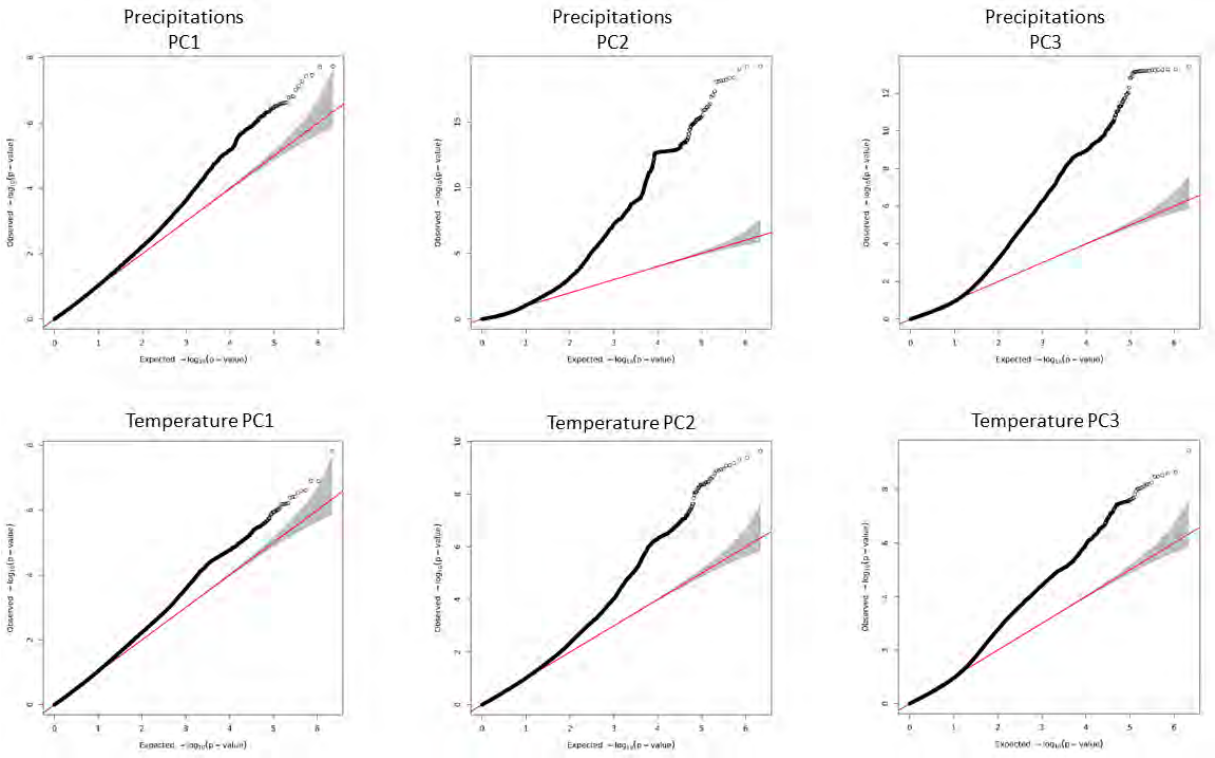




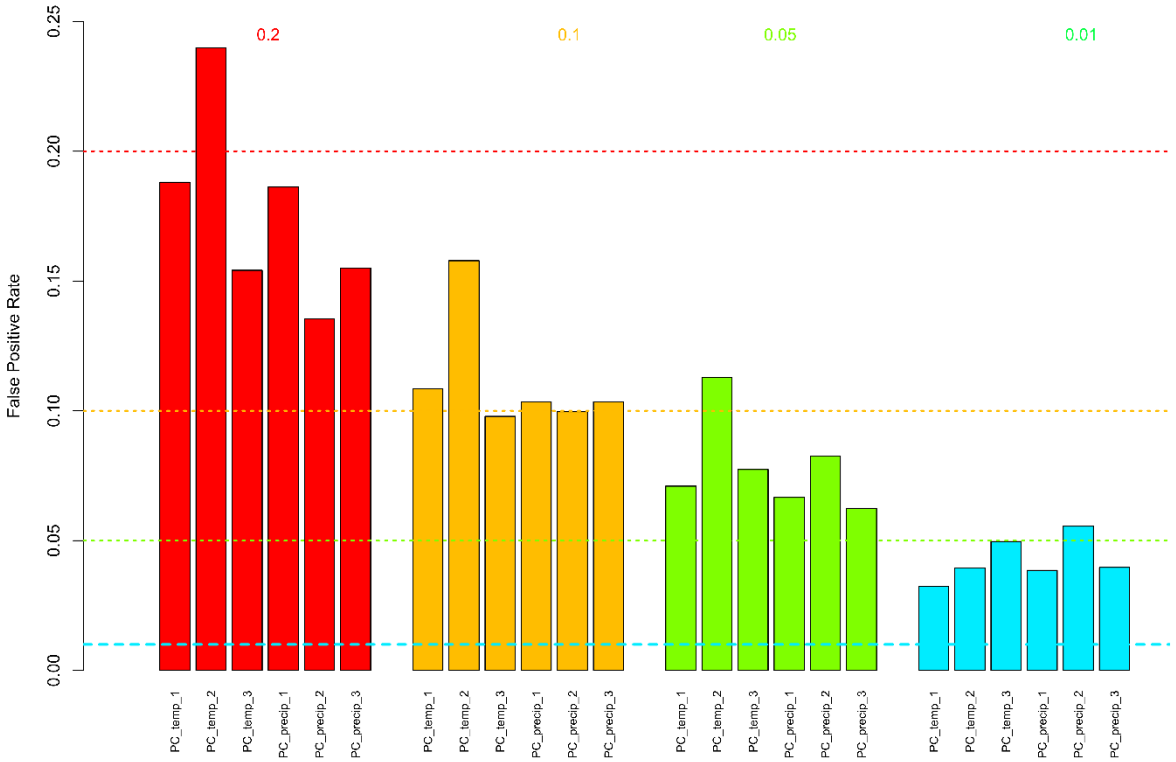


Appendix A: Linkage disequilibrium decay in three species: *Marchantia polymorpha*, *Arabidopsis thaliana* and *Medicago truncatula*. The half LD decay is respectively of 3.6 kb, 0.9 kb and 0.6 kb

Appendix B: QQplots showing the distribution of p-values GEA for the 3 first PC of the precipitation and temperature linked variables.



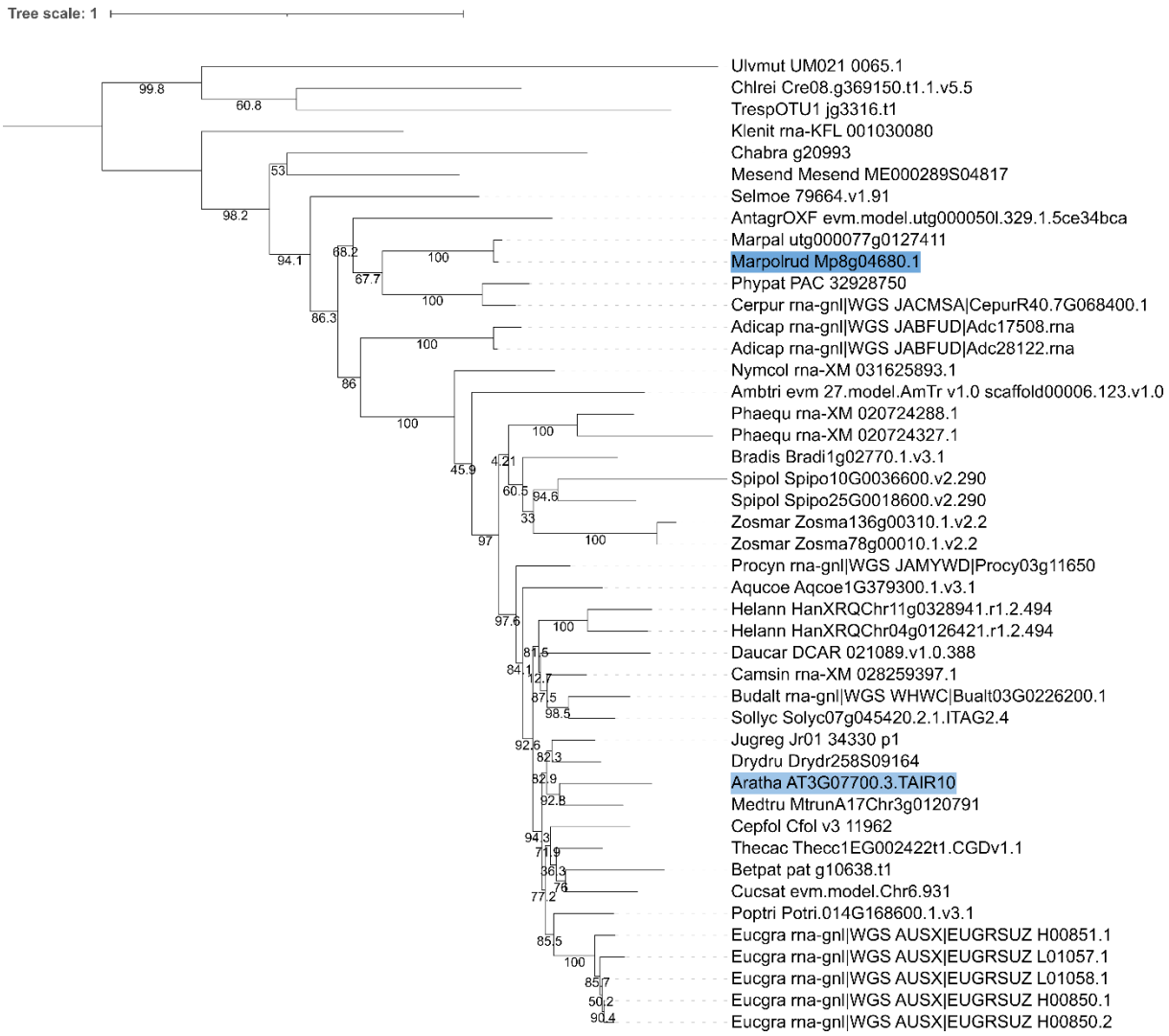
Appendix C: number of SNPs in the top 20% (red), 10% (orange), 5% (green) or 1% (blue) quantiles of SNPs with the lowest p-values. Each bar represents data from a genome wide association study on climatic data (the first 3 PC of temperature and precipitation are represented). The horizontal lines represent the expected number of SNPs if the p-values followed the theoretical distribution. P-values from most climatic data are below or equal the expected p-values for the 20 and 10% quantile of smallest p-values, and then above the 5% and 1% quantile. This means that only the top 5% of SNPs with the smallest p-values are more significant than expected, and therefore that there must not be too many false positives, but mostly SNPs truly associated with the climatic data. Only the SNP p-values associated to the second PC on precipitations are above what is expected even in the 20% quantile which confirms the trend observed on the QQplots, with p-values way higher than what was expected for a great number of SNPs.



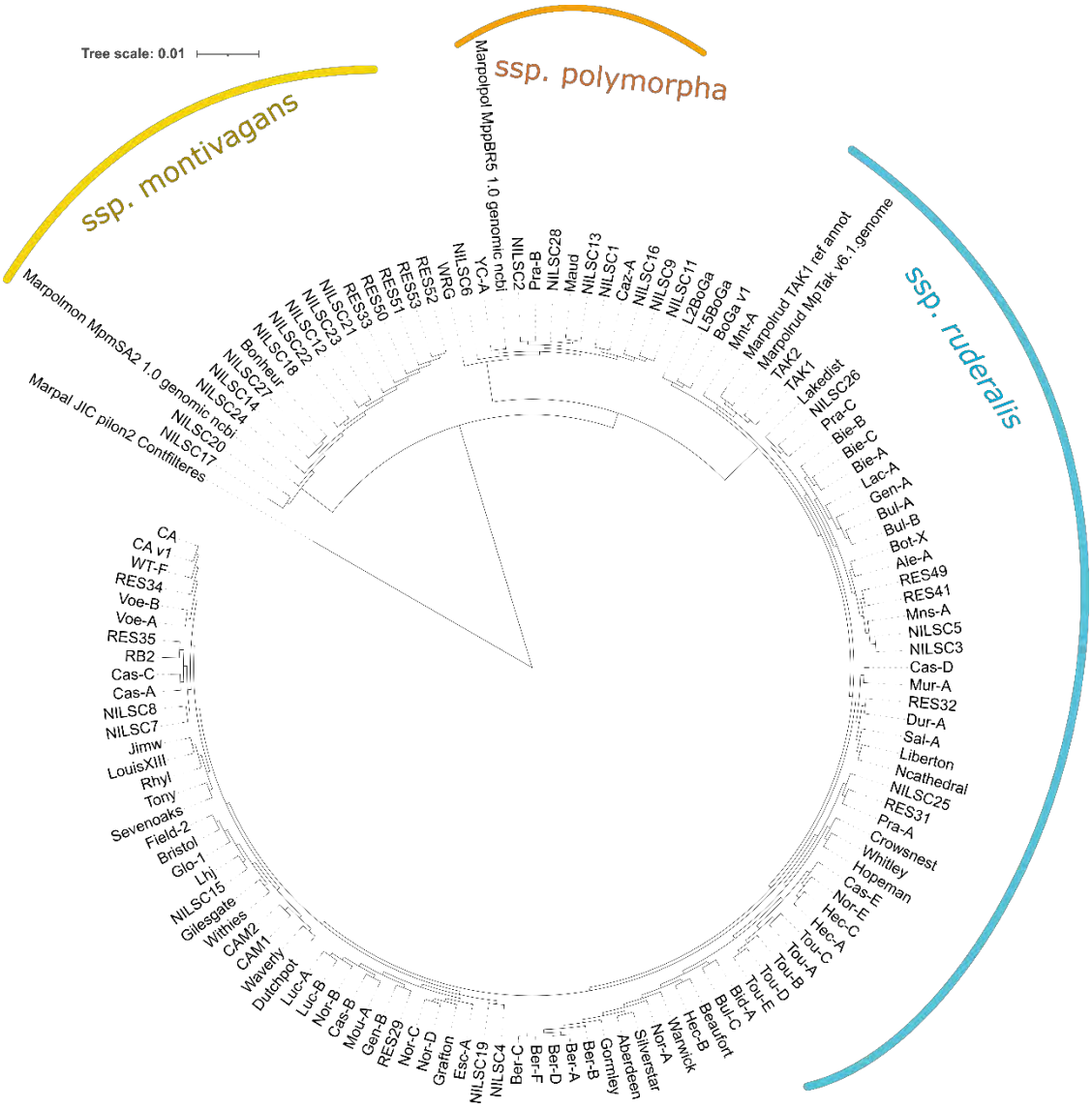
Appendix D: Table of the 37 representative Viridiplantae species used in phylogenies of the GEA and GWAS candidates

Species Code	Species Name	Species Group
Ambtri	*Amborella trichopoda*	Angiosperms
Aqucoe	*Aquilegia coerulea*	Angiosperms
Aratha	*Arabidopsis thaliana*	Angiosperms
Betpat	*Beta patula*	Angiosperms
Bradis	*Brachypodium distachyon*	Angiosperms
Budalt	*Buddleja alternifolia*	Angiosperms
Camsin	*Camellia sinensis*	Angiosperms
Cepfol	*Cephalotus follicularis*	Angiosperms
Cucsat	*Cucumis sativus PI183967*	Angiosperms
Daucar	*Daucus carota*	Angiosperms
Drydru	*Dryas drummondii*	Angiosperms
Eucgra	*Eucalyptus grandis*	Angiosperms
Helann	*Helianthus annuus*	Angiosperms
Jugreg	*Juglans regia*	Angiosperms
Medtru	*Medicago truncatula*	Angiosperms
Nymcol	*Nymphea colarata*	Angiosperms
Phaequ	*Phalaenopsis equestris*	Angiosperms
Procyn	*Protea cynaroides*	Angiosperms
Sollyc	*Solanum lycopersicum*	Angiosperms
Spipol	*Spirodela polyrhiza*	Angiosperms
Thecac	*Theobrama cacao*	Angiosperms
Zosmar	*Zostera marina*	Angiosperms
Chlrei	*Chlamydomonas reinhardtii*	Chlorophytes
TrespOTU1	*Trebouxia sp. OTU1 generalist lineage*	Chlorophytes
Ulvmut	*Ulva mutabilis*	Chlorophytes
Cycpan	*Cycas panzhihuaensis*	Gymnosperms
AntagrOXF	*Anthoceros agrestis cv. OXF*	Hornworts
Marpolrud	*Marchantia polymorpha ssp. ruderalis TAK1 v6.1*	Liverworts
Selmoe	*Selaginella moellendorffii*	Lycophytes
Adicap	*Adiantum capillus-veneris*	Monilophytes
Cerpur	*Ceratodon purpureus*	Mosses
Phypat	*Physcomitrium patens*	Mosses
Marpal	*Marchantia paleaceae*	Liverworts
Poptri	*Populus trichocarpa*	Angiosperms
Mesend	*Mesotaenium endlicherianum*	Charophytes
Chabra	*Chara braunii*	Charophytes
Klenit	*Klebsormidium nitens*	Charophytes

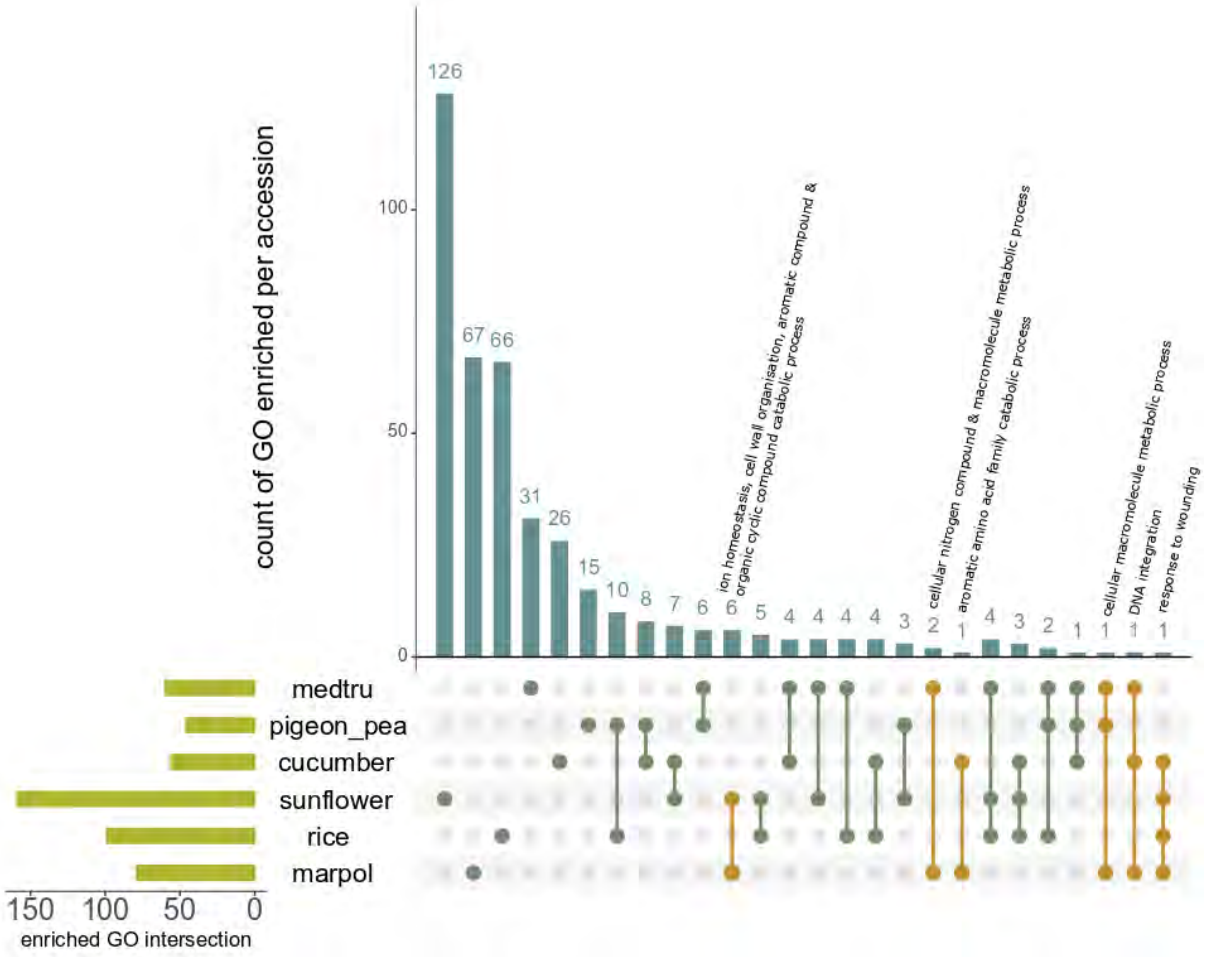
Appendix E: Phylogenetic tree of the orthologs of the ABC1 gene of *M. polymorpha* ssp. *ruderalis* illustrating a direct orthology with the ABC1K7 gene in *A. thaliana*. This tree was computed with the substitution model Q.plant+R5 and has a log-likelihood of -26208.0663



Appendix F: Phylogenetic tree of the 134 genomes used in the pangenome (accessions from the 3 subspecies of *M. polymorpha* and a *M. paleacea* genome). This species tree results from the reconciliation of the gene trees by the Orthofinder analysis. All accession were correctly assigned to their subspecies, and the rooting with *M. paleacea* gives a similar topology between the three subspecies that the one gotten by **(Linde et al., 2020)**: the *ruderalis* and *polymorpha* subspecies are sisters, and the *montivagans* subspecies is a bit more distant.



Appendix G: UpSet plot of the comparison of GO term enrichment of accessory genome in 6 different plant pangenomes. No GO term was found enriched in all pangenomes because of the heterogeneity of the pangenomic data.



Bibliography

Bibliography

Abd-Hamid, N.-A., Ahmad-Fauzi, M.-I., Zainal, Z., & Ismail, I. (2020). Diverse and dynamic roles of F-box proteins in plant biology. *Planta*, *251*(3), 68. https://doi.org/10.1007/s00425-020-03356-8

Alcaraz, L. D., Peimbert, M., Barajas, H. R., Dorantes-Acosta, A. E., Bowman, J. L., & Arteaga-Vázquez, M. A. (2018). Marchantia liverworts as a proxy to plants' basal microbiomes. *Scientific Reports*, *8*(1), 12712. https://doi.org/10.1038/s41598-018-31168-0

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*(9), 1655-1664. https://doi.org/10.1101/gr.094052.109

Ali, S., Ganai, B. A., Kamili, A. N., Bhat, A. A., Mir, Z. A., Bhat, J. A., Tyagi, A., Islam, S. T., Mushtaq, M., Yadav, P., Rawat, S., & Grover, A. (2018). Pathogenesis-related proteins and peptides as promising tools for engineering plants with multiple stress tolerance. *Microbiological Research*, *212-213*, 29-37. https://doi.org/10.1016/j.micres.2018.04.008

Almagro, L., Gómez Ros, L. V., Belchi-Navarro, S., Bru, R., Ros Barceló, A., & Pedreño, M. A. (2009). Class III peroxidases in plant defence reactions. *Journal of Experimental Botany*, *60*(2), 377-390. https://doi.org/10.1093/jxb/ern277

Alonso-Blanco, C., Andrade, J., Becker, C., Bemm, F., Bergelson, J., Borgwardt, K. M., Cao, J., Chae, E., Dezwaan, T. M., Ding, W., Ecker, J. R., Exposito-Alonso, M., Farlow, A., Fitz, J., Gan, X., Grimm, D. G., Hancock, A. M., Henz, S. R., Holm, S., … Zhou, X. (2016). 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell*, *166*(2), 481-491. https://doi.org/10.1016/j.cell.2016.05.063

Alseekh, S., Kostova, D., Bulut, M., & Fernie, A. R. (2021). Genome-wide association studies : Assessing trait characteristics in model and crop plants. *Cellular and Molecular Life Sciences*, *78*(15), 5743-5754. https://doi.org/10.1007/s00018-021-03868-w

Andolfo, G., Donato, A. D., Chiaiese, P., Natale, A. D., Pollio, A., Jones, J. D. G., Frusciante, L., & Ercolano, M. R. (2019). Alien domains shaped the modular structure of plant NLR proteins. *Genome Biology and Evolution*, evz248. https://doi.org/10.1093/gbe/evz248

Aranzana, M. J., Kim, S., Zhao, K., Bakker, E., Horton, M., Jakob, K., Lister, C., Molitor, J., Shindo, C., Tang, C., Toomajian, C., Traw, B., Zheng, H., Bergelson, J., Dean, C., Marjoram, P., & Nordborg, M. (2005). Genome-Wide Association Mapping in Arabidopsis Identifies Previously Known Flowering Time and Pathogen Resistance Genes. *PLoS Genetics*, *1*(5), e60. https://doi.org/10.1371/journal.pgen.0010060

Ariga, H., Katori, T., Tsuchimatsu, T., Hirase, T., Tajima, Y., Parker, J. E., Alcázar, R., Koornneef, M., Hoekenga, O., Lipka, A. E., Gore, M. A., Sakakibara, H., Kojima, M., Kobayashi, Y., Iuchi, S., Kobayashi, M., Shinozaki, K., Sakata, Y., Hayashi, T., … Taji, T. (2017). NLR locus-mediated trade-off between abiotic and biotic stress adaptation in Arabidopsis. *Nature Plants*, *3*(6), 17072. https://doi.org/10.1038/nplants.2017.72

Asakawa, Y., & Ludwiczuk, A. (2018). Chemical Constituents of Bryophytes : Structures and Biological Activity. *Journal of Natural Products*, *81*(3), 641-660. https://doi.org/10.1021/acs.jnatprod.6b01046

Baaijens, J. A., Bonizzoni, P., Boucher, C., Della Vedova, G., Pirola, Y., Rizzi, R., & Sirén, J. (2022). Computational graph pangenomics : A tutorial on data structures and their applications. *Natural Computing*, *21*(1), 81-108. https://doi.org/10.1007/s11047-022-09882-6

Baggs, E. L., Monroe, J. G., Thanki, A. S., O'Grady, R., Schudoma, C., Haerty, W., & Krasileva, K. V. (2020). Convergent Loss of an EDS1/PAD4 Signaling Pathway in Several Plant Lineages Reveals Coevolved Components of Plant Immunity and Drought Response. *The Plant Cell*, *32*(7), 2158-2177. https://doi.org/10.1105/tpc.19.00903

Baig, A. (2018). Role of Arabidopsis LOR1 (URP-one related one) in basal defense against Hyaloperonospora arabidopsidis. *Physiological and Molecular Plant Pathology*, *103*, 71-77. https://doi.org/10.1016/j.pmpp.2018.05.003

Barragan, A. C., & Weigel, D. (2021). Plant NLR diversity : The known unknowns of pan-NLRomes. *The Plant Cell*, *33*(4), 814-831. https://doi.org/10.1093/plcell/koaa002

Bayer, P. E., Golicz, A. A., Scheben, A., Batley, J., & Edwards, D. (2020). Plant pan-genomes are the new reference. *Nature Plants*, *6*(8), 914-920. https://doi.org/10.1038/s41477-020-0733-0

Beaumont, M. A. (2005). Adaptation and speciation : What can Fst tell us? *Trends in Ecology & Evolution*, *20*(8), 435-440. https://doi.org/10.1016/j.tree.2005.05.017

Beerli, P., & Felsenstein, J. (1999). Maximum-Likelihood Estimation of Migration Rates and Effective Population Numbers in Two Populations Using a Coalescent Approach. *Genetics*, *152*(2), 763-773. https://doi.org/10.1093/genetics/152.2.763

Benton, M. J., Wilf, P., & Sauquet, H. (2022). The Angiosperm Terrestrial Revolution and the origins of modern biodiversity. *New Phytologist*, *233*(5), 2017-2035. https://doi.org/10.1111/nph.17822

Berry, A., & Browne, J. (2022). Mendel and Darwin. *Proceedings of the National Academy of Sciences*, *119*(30), e2122144119. https://doi.org/10.1073/pnas.2122144119

Bohra, A., Kilian, B., Sivasankar, S., Caccamo, M., Mba, C., McCouch, S. R., & Varshney, R. K. (2022). Reap the crop wild relatives for breeding future crops. *Trends in Biotechnology*, *40*(4), 412-431. https://doi.org/10.1016/j.tibtech.2021.08.009

Bonhomme, M., Chevalet, C., Servin, B., Boitard, S., Abdallah, J., Blott, S., & SanCristobal, M. (2010). Detecting Selection in Population Trees : The Lewontin and Krakauer Test Extended. *Genetics*, *186*(1), 241-262. https://doi.org/10.1534/genetics.110.117275

Bonhomme, M., Fariello, M. I., Navier, H., Hajri, A., Badis, Y., Miteul, H., Samac, D. A., Dumas, B., Baranger, A., Jacquet, C., & Pilet-Nayel, M.-L. (2019). A local score approach improves GWAS resolution and detects minor QTL : Application to Medicago truncatula quantitative disease resistance to multiple Aphanomyces euteiches isolates. *Heredity*, *123*(4), 517-531. https://doi.org/10.1038/s41437-019-0235-x

Bonhomme, M., & Jacquet, C. (2020). Genome-wide association mapping and population genomic features in *Medicago truncatula*. In F. De Bruijn (Éd.), *The Model Legume Medicago truncatula* (1re éd., p. 870-881). Wiley. https://doi.org/10.1002/9781119409144.ch109

Bowman, J. L. (2016). A Brief History of Marchantia from Greece to Genomics. *Plant and Cell Physiology*, *57*(2), 210-229. https://doi.org/10.1093/pcp/pcv044

Bowman, J. L. (2022). The origin of a land flora. *Nature Plants*, *8*(12), 1352-1369. https://doi.org/10.1038/s41477-022-01283-y

Bowman, J. L., Arteaga-Vazquez, M., Berger, F., Briginshaw, L. N., Carella, P., Aguilar-Cruz, A., Davies, K. M., Dierschke, T., Dolan, L., Dorantes-Acosta, A. E., Fisher, T. J., Flores-Sandoval, E., Futagami, K.,

Ishizaki, K., Jibran, R., Kanazawa, T., Kato, H., Kohchi, T., Levins, J., … Zachgo, S. (2022). The renaissance and enlightenment of *Marchantia* as a model system. *The Plant Cell*, *34*(10), 3512-3542. https://doi.org/10.1093/plcell/koac219

Bowman, J. L., Kohchi, T., Yamato, K. T., Jenkins, J., Shu, S., Ishizaki, K., Yamaoka, S., Nishihama, R., Nakamura, Y., Berger, F., Adam, C., Aki, S. S., Althoff, F., Araki, T., Arteaga-Vazquez, M. A., Balasubrmanian, S., Barry, K., Bauer, D., Boehm, C. R., … Schmutz, J. (2017a). Insights into Land Plant Evolution Garnered from the <i>Marchantia polymorpha<i> Genome. *Cell*, *171*(2), 287-304.e15. https://doi.org/10.1016/j.cell.2017.09.030

Bowman, J. L., Kohchi, T., Yamato, K. T., Jenkins, J., Shu, S., Ishizaki, K., Yamaoka, S., Nishihama, R., Nakamura, Y., Berger, F., Adam, C., Aki, S. S., Althoff, F., Araki, T., Arteaga-Vazquez, M. A., Balasubrmanian, S., Barry, K., Bauer, D., Boehm, C. R., … Schmutz, J. (2017b). Insights into Land Plant Evolution Garnered from the Marchantia polymorpha Genome. *Cell*, *171*(2), 287-304.e15. https://doi.org/10.1016/j.cell.2017.09.030

Boyes, D. C., Nam, J., & Dangl, J. L. (1998). The *Arabidopsis thaliana RPM1* disease resistance gene product is a peripheral plasma membrane protein that is degraded coincident with the hypersensitive response. *Proceedings of the National Academy of Sciences*, *95*(26), 15849-15854. https://doi.org/10.1073/pnas.95.26.15849

Boyle, E. A., Li, Y. I., & Pritchard, J. K. (2017). An Expanded View of Complex Traits : From Polygenic to Omnigenic. *Cell*, *169*(7), 1177-1186. https://doi.org/10.1016/j.cell.2017.05.038

Breen, S., Williams, S. J., Outram, M., Kobe, B., & Solomon, P. S. (2017). Emerging Insights into the Functions of Pathogenesis-Related Protein 1. *Trends in Plant Science*, *22*(10), 871-879. https://doi.org/10.1016/j.tplants.2017.06.013

Brůna, T., Hoff, K. J., Lomsadze, A., Stanke, M., & Borodovsky, M. (2021). BRAKER2 : Automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genomics and Bioinformatics*, *3*(1), lqaa108. https://doi.org/10.1093/nargab/lqaa108

Buchfink, B., Reuter, K., & Drost, H.-G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods*, *18*(4), 366-368. https://doi.org/10.1038/s41592-021-01101-x

Burgarella, C., Chantret, N., Gay, L., Prosperi, J., Bonhomme, M., Tiffin, P., Young, N. D., & Ronfort, J. (2016). Adaptation to climate through flowering phenology : A case study in *Medicago truncatula*. *Molecular Ecology*, *25*(14), 3397-3415. https://doi.org/10.1111/mec.13683

Burgeff, H. (1943). *Genetische Studien an Marchantia : Einführung einer neuen Pflanzenfamilie in die genetische Wissenschaft.*

Cabanettes, F., & Klopp, C. (2018). D-GENIES : Dot plot large genomes in an interactive, efficient and simple way. *PeerJ*, *6*, e4958. https://doi.org/10.7717/peerj.4958

Carella, P., Gogleva, A., Hoey, D. J., Bridgen, A. J., Stolze, S. C., Nakagami, H., & Schornack, S. (2019). Conserved Biochemical Defenses Underpin Host Responses to Oomycete Infection in an Early-Divergent Land Plant Lineage. *Current Biology*, *29*(14), 2282-2294.e5. https://doi.org/10.1016/j.cub.2019.05.078

Carella, P., Gogleva, A., Tomaselli, M., Alfs, C., & Schornack, S. (2018). *Phytophthora palmivora* establishes tissue-specific intracellular infection structures in the earliest divergent land plant lineage. *Proceedings of the National Academy of Sciences*, *115*(16). https://doi.org/10.1073/pnas.1717900115

Carey, S. B., Jenkins, J., Lovell, J. T., Maumus, F., Sreedasyam, A., Payton, A. C., Shu, S., Tiley, G. P., Fernandez-Pozo, N., Healey, A., Barry, K., Chen, C., Wang, M., Lipzen, A., Daum, C., Saski, C. A., McBreen, J. C., Conrad, R. E., Kollar, L. M., … McDaniel, S. F. (2021). Gene-rich UV sex chromosomes harbor conserved regulators of sexual development. *Science Advances*, *7*(27), eabh2488. https://doi.org/10.1126/sciadv.abh2488

Cesarino, I., Dello Ioio, R., Kirschner, G. K., Ogden, M. S., Picard, K. L., Rast-Somssich, M. I., & Somssich, M. (2020). Plant science's next top models. *Annals of Botany*, *126*(1), 1-23. https://doi.org/10.1093/aob/mcaa063

Chang, C., Bowman, J. L., & Meyerowitz, E. M. (2016). Field Guide to Plant Model Systems. *Cell*, *167*(2), 325-339. https://doi.org/10.1016/j.cell.2016.08.031

Charlesworth, B., & Charlesworth, D. (2017). Population genetics from 1966 to 2016. *Heredity*, *118*(1), 2-9. https://doi.org/10.1038/hdy.2016.55

Charlesworth, D. (2006). Balancing Selection and Its Effects on Sequences in Nearby Genome Regions. *PLoS Genetics*, *2*(4), e64. https://doi.org/10.1371/journal.pgen.0020064

Charlesworth, D., & Wright, S. I. (2001). Breeding systems and genome evolution. *Current Opinion in Genetics & Development*, *11*(6), 685-690. https://doi.org/10.1016/S0959-437X(00)00254-9

Chen, F., Ludwiczuk, A., Wei, G., Chen, X., Crandall-Stotler, B., & Bowman, J. L. (2018). Terpenoid Secondary Metabolites in Bryophytes : Chemical Diversity, Biosynthesis and Biological Functions. *Critical Reviews in Plant Sciences*, *37*(2-3), 210-231. https://doi.org/10.1080/07352689.2018.1482397

Chen, H., He, C., Wang, C., Wang, X., Ruan, F., Yan, J., Yin, P., Wang, Y., & Yan, S. (2021). RAD51 supports DMC1 by inhibiting the SMC5/6 complex during meiosis. *The Plant Cell*, *33*(8), 2869-2882. https://doi.org/10.1093/plcell/koab136

Cheng, F., Wu, J., Cai, X., Liang, J., Freeling, M., & Wang, X. (2018). Gene retention, fractionation and subgenome differences in polyploid plants. *Nature Plants*, *4*(5), 258-268. https://doi.org/10.1038/s41477-018-0136-7

Cheng, X., Xiong, R., Yan, H., Gao, Y., Liu, H., Wu, M., & Xiang, Y. (2019). The trihelix family of transcription factors : Functional and evolutionary analysis in Moso bamboo (Phyllostachys edulis). *BMC Plant Biology*, *19*(1), 154. https://doi.org/10.1186/s12870-019-1744-8

Chia, K., & Carella, P. (2023). Taking the lead : NLR immune receptor N-terminal domains execute plant immune responses. *New Phytologist*, *240*(2), 496-501. https://doi.org/10.1111/nph.19170

Chia, K.-S., Kourelis, J., Vickers, M., Sakai, T., Kamoun, S., & Carella, P. (2022). *The N-terminal executioner domains of NLR immune receptors are functionally conserved across major plant lineages* [Preprint]. Plant Biology. https://doi.org/10.1101/2022.10.19.512840

Chini, A., Grant, J. J., Seki, M., Shinozaki, K., & Loake, G. J. (2004). Drought tolerance established by enhanced expression of the *CC–NBS–LRR* gene, *ADR1* , requires salicylic acid, EDS1 and ABI1. *The Plant Journal*, *38*(5), 810-822. https://doi.org/10.1111/j.1365-313X.2004.02086.x

Christenhusz, M. J. M., & Byng, J. W. (2016). The number of known plants species in the world and its annual increase. *Phytotaxa*, *261*(3), 201. https://doi.org/10.11646/phytotaxa.261.3.1

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms,

SnpEff : SNPs in the genome of Drosophila melanogaster strain w ^{1118} ; iso-2; iso-3. *Fly*, *6*(2), 80‑92. https://doi.org/10.4161/fly.19695

Clark, J. W., & Donoghue, P. C. J. (2018). Whole-Genome Duplication and Plant Macroevolution. *Trends in Plant Science*, *23*(10), 933‑945. https://doi.org/10.1016/j.tplants.2018.07.006

Cook, D. E., Lee, T. G., Guo, X., Melito, S., Wang, K., Bayless, A. M., Wang, J., Hughes, T. J., Willis, D. K., Clemente, T. E., Diers, B. W., Jiang, J., Hudson, M. E., & Bent, A. F. (2012). Copy Number Variation of Multiple Genes at *Rhg1* Mediates Nematode Resistance in Soybean. *Science*, *338*(6111), 1206‑1209. https://doi.org/10.1126/science.1228746

Cortés, A. J., López-Hernández, F., & Blair, M. W. (2022). Genome–Environment Associations, an Innovative Tool for Studying Heritable Evolutionary Adaptation in Orphan Crops and Wild Relatives. *Frontiers in Genetics*, *13*, 910386. https://doi.org/10.3389/fgene.2022.910386

Crow, J. F., & Kimura, M. (2009). *An introduction to Population Genetics Theory*. Blackburn Press.

Croze, M., Živković, D., Stephan, W., & Hutter, S. (2016). Balancing selection on immunity genes : Review of the current literature and new analysis in Drosophila melanogaster. *Zoology*, *119*(4), 322‑329. https://doi.org/10.1016/j.zool.2016.03.004

Cubry, P., Tranchant-Dubreuil, C., Thuillet, A.-C., Monat, C., Ndjiondjop, M.-N., Labadie, K., Cruaud, C., Engelen, S., Scarcelli, N., Rhoné, B., Burgarella, C., Dupuy, C., Larmande, P., Wincker, P., François, O., Sabot, F., & Vigouroux, Y. (2018). The Rise and Fall of African Rice Cultivation Revealed by Analysis of 246 New Genomes. *Current Biology*, *28*(14), 2274-2282.e6. https://doi.org/10.1016/j.cub.2018.05.066

Dao, T. T. H., Linthorst, H. J. M., & Verpoorte, R. (2011). Chalcone synthase and its functions in plant resistance. *Phytochemistry Reviews*, *10*(3), 397‑412. https://doi.org/10.1007/s11101-011-9211-7

Darwin, C. (2009). *The Origin of Species : By Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life* (6^{e} éd.). Cambridge University Press. https://doi.org/10.1017/CBO9780511694295

De Bruyne, L., Höfte, M., & De Vleesschauwer, D. (2014). Connecting Growth and Defense : The Emerging Roles of Brassinosteroids and Gibberellins in Plant Innate Immunity. *Molecular Plant*, *7*(6), 943‑959. https://doi.org/10.1093/mp/ssu050

De Coninck, T., & Van Damme, E. J. M. (2021). Review : The multiple roles of plant lectins. *Plant Science*, *313*, 111096. https://doi.org/10.1016/j.plantsci.2021.111096

Delaux, P.-M., Hetherington, A. J., Coudert, Y., Delwiche, C., Dunand, C., Gould, S., Kenrick, P., Li, F.-W., Philippe, H., Rensing, S. A., Rich, M., Strullu-Derrien, C., & de Vries, J. (2019). Reconstructing trait evolution in plant evo–devo studies. *Current Biology*, *29*(21), R1110‑R1118. https://doi.org/10.1016/j.cub.2019.09.044

Delaux, P.-M., & Schornack, S. (2021). Plant evolution driven by interactions with symbiotic and pathogenic microbes. *Science*, *371*(6531), eaba6605. https://doi.org/10.1126/science.aba6605

Della Coletta, R., Qiu, Y., Ou, S., Hufford, M. B., & Hirsch, C. N. (2021). How the pan-genome is changing crop genomics and improvement. *Genome Biology*, *22*(1), 3. https://doi.org/10.1186/s13059-020-02224-8

Delwiche, C. F., & Cooper, E. D. (2015). The Evolutionary Origin of a Terrestrial Flora. *Current Biology*, *25*(19), R899‑R910. https://doi.org/10.1016/j.cub.2015.08.029

Demirjian, C., Vailleau, F., Berthomé, R., & Roux, F. (2023). Genome-wide association studies in plant pathosystems : Success or failure? *Trends in Plant Science*, *28*(4), 471-485. https://doi.org/10.1016/j.tplants.2022.11.006

de Vries, J., & Ischebeck, T. (2020). Ties between Stress and Lipid Droplets Pre-date Seeds. *Trends in Plant Science*, *25*(12), 1203-1214. https://doi.org/10.1016/j.tplants.2020.07.017

Didelon, M., Khafif, M., Godiard, L., Barbacci, A., & Raffaele, S. (2020). Patterns of Sequence and Expression Diversification Associate Members of the PADRE Gene Family With Response to Fungal Pathogens. *Frontiers in Genetics*, *11*, 491. https://doi.org/10.3389/fgene.2020.00491

Diop, S. I., Subotic, O., Giraldo-Fonseca, A., Waller, M., Kirbis, A., Neubauer, A., Potente, G., Murray-Watson, R., Boskovic, F., Bont, Z., Hock, Z., Payton, A. C., Duijsings, D., Pirovano, W., Conti, E., Grossniklaus, U., McDaniel, S. F., & Szövényi, P. (2020). A pseudomolecule-scale genome assembly of the liverwort <i>Marchantia polymorpha<i>. *The Plant Journal*, *101*(6), 1378-1396. https://doi.org/10.1111/tpj.14602

Donoghue, P. C. J., Harrison, C. J., Paps, J., & Schneider, H. (2021). The evolutionary emergence of land plants. *Current Biology*, *31*(19), R1281-R1298. https://doi.org/10.1016/j.cub.2021.07.038

Dronamraju, K. (2015). J.B.S. Haldane as I knew him, with a brief account of his contribution to mutation research. *Mutation Research/Reviews in Mutation Research*, *765*, 1-6. https://doi.org/10.1016/j.mrrev.2015.05.002

Dussex, N., Van Der Valk, T., Morales, H. E., Wheat, C. W., Díez-del-Molino, D., Von Seth, J., Foster, Y., Kutschera, V. E., Guschanski, K., Rhie, A., Phillippy, A. M., Korlach, J., Howe, K., Chow, W., Pelan, S., Mendes Damas, J. D., Lewin, H. A., Hastie, A. R., Formenti, G., … Dalén, L. (2021). Population genomics of the critically endangered kākāpō. *Cell Genomics*, *1*(1), 100002. https://doi.org/10.1016/j.xgen.2021.100002

Eckardt, N. A., Ainsworth, E. A., Bahuguna, R. N., Broadley, M. R., Busch, W., Carpita, N. C., Castrillo, G., Chory, J., DeHaan, L. R., Duarte, C. M., Henry, A., Jagadish, S. V. K., Langdale, J. A., Leakey, A. D. B., Liao, J. C., Lu, K.-J., McCann, M. C., McKay, J. K., Odeny, D. A., … Zhang, X. (2023). Climate change challenges, plant science solutions. *The Plant Cell*, *35*(1), 24-66. https://doi.org/10.1093/plcell/koac303

Emms, D. M., & Kelly, S. (2019). OrthoFinder : Phylogenetic orthology inference for comparative genomics. *Genome Biology*, *20*(1), 238. https://doi.org/10.1186/s13059-019-1832-y

Epstein, B., Burghardt, L. T., Heath, K. D., Grillo, M. A., Kostanecki, A., Hämälä, T., Young, N. D., & Tiffin, P. (2022). Combining GWAS and population genomic analyses to characterize coevolution in a legume-rhizobia symbiosis. *Molecular Ecology*, mec.16602. https://doi.org/10.1111/mec.16602

Espiñeira, J. M., Novo Uzal, E., Gómez Ros, L. V., Carrión, J. S., Merino, F., Ros Barceló, A., & Pomar, F. (2011). Distribution of lignin monomers and the evolution of lignification among lower plants. *Plant Biology*, *13*(1), 59-68. https://doi.org/10.1111/j.1438-8677.2010.00345.x

Fariello, M. I., Boitard, S., Mercier, S., Robelin, D., Faraut, T., Arnould, C., Recoquillay, J., Bouchez, O., Salin, G., Dehais, P., Gourichon, D., Leroux, S., Pitel, F., Leterrier, C., & SanCristobal, M. (2017). Accounting for linkage disequilibrium in genome scans for selection without individual genotypes : The local score approach. *Molecular Ecology*, *26*(14), 3700-3714. https://doi.org/10.1111/mec.14141

Fay, J. C., & Wu, C. I. (2000). Hitchhiking under positive Darwinian selection. *Genetics*, *155*(3), 1405-1413.

Ferrero-Serrano, Á., & Assmann, S. M. (2019). Phenotypic and genome-wide association with the local environment of Arabidopsis. *Nature Ecology & Evolution*, *3*(2), 274-285. https://doi.org/10.1038/s41559-018-0754-5

Ferretti, L., Raineri, E., & Ramos-Onsins, S. (2012). Neutrality Tests for Sequences with Missing Data. *Genetics*, *191*(4), 1397-1401. https://doi.org/10.1534/genetics.112.139949

Feussner, I., & Wasternack, C. (2002). The lipoxygenase pathway. *Annual Review of Plant Biology*, *53*(1), 275-297. https://doi.org/10.1146/annurev.arplant.53.100301.135248

Fick, S. E., & Hijmans, R. J. (2017). WorldClim 2 : New 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, *37*(12), 4302-4315. https://doi.org/10.1002/joc.5086

Frachon, L., Bartoli, C., Carrère, S., Bouchez, O., Chaubet, A., Gautier, M., Roby, D., & Roux, F. (2018). A Genomic Map of Climate Adaptation in Arabidopsis thaliana at a Micro-Geographic Scale. *Frontiers in Plant Science*, *9*, 967. https://doi.org/10.3389/fpls.2018.00967

Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT : Accelerated for clustering the next-generation sequencing data. *Bioinformatics*, *28*(23), 3150-3152. https://doi.org/10.1093/bioinformatics/bts565

Fukao, T., Harris, T., & Bailey-Serres, J. (2009). Evolutionary analysis of the Sub1 gene cluster that confers submergence tolerance to domesticated rice. *Annals of Botany*, *103*(2), 143-150. https://doi.org/10.1093/aob/mcn172

Fürst-Jansen, J. M. R., De Vries, S., & De Vries, J. (2020). Evo-physio : On stress responses and the earliest land plants. *Journal of Experimental Botany*, *71*(11), 3254-3269. https://doi.org/10.1093/jxb/eraa007

Gage, J. L., Vaillancourt, B., Hamilton, J. P., Manrique-Carpintero, N. C., Gustafson, T. J., Barry, K., Lipzen, A., Tracy, W. F., Mikel, M. A., Kaeppler, S. M., Buell, C. R., & De Leon, N. (2019). Multiple Maize Reference Genomes Impact the Identification of Variants by Genome-Wide Association Study in a Diverse Inbred Panel. *The Plant Genome*, *12*(2), 180069. https://doi.org/10.3835/plantgenome2018.09.0069

Galindo-Trigo, S., Gray, J. E., & Smith, L. M. (2016). Conserved Roles of CrRLK1L Receptor-Like Kinases in Cell Expansion and Reproduction from Algae to Angiosperms. *Frontiers in Plant Science*, *07*. https://doi.org/10.3389/fpls.2016.01269

Gao, B., Chen, M., Li, X., Liang, Y., Zhang, D., Wood, A. J., Oliver, M. J., & Zhang, J. (2022). Ancestral gene duplications in mosses characterized by integrated phylogenomic analyses. *Journal of Systematics and Evolution*, *60*(1), 144-159. https://doi.org/10.1111/jse.12683

Gao, Y.-Q., Huang, J.-Q., Reyt, G., Song, T., Love, A., Tiemessen, D., Xue, P.-Y., Wu, W.-K., George, M. W., Chen, X.-Y., Chao, D.-Y., Castrillo, G., & Salt, D. E. (2023). A dirigent protein complex directs lignin polymerization and assembly of the root diffusion barrier. *Science*, *382*(6669), 464-471. https://doi.org/10.1126/science.adi5032

Garrison, E., Guarracino, A., Heumos, S., Villani, F., Bao, Z., Tattini, L., Hagmann, J., Vorbrugg, S., Marco-Sola, S., Kubica, C., Ashbrook, D. G., Thorell, K., Rusholme-Pilcher, R. L., Liti, G., Rudbeck, E.,

Nahnsen, S., Yang, Z., Moses, M. N., Nobrega, F. L., … Prins, P. (2023). *Building pangenome graphs* [Preprint]. Bioinformatics. https://doi.org/10.1101/2023.04.05.535718

Gharabli, H., Della Gala, V., & Welner, D. H. (2023). The function of UDP-glycosyltransferases in plants and their possible use in crop protection. *Biotechnology Advances*, *67*, 108182. https://doi.org/10.1016/j.biotechadv.2023.108182

Gimenez-Ibanez, S., Zamarreño, A. M., García-Mina, J. M., & Solano, R. (2019). An Evolutionarily Ancient Immune System Governs the Interactions between Pseudomonas syringae and an Early-Diverging Land Plant Lineage. *Current Biology*, *29*(14), 2270-2281.e4. https://doi.org/10.1016/j.cub.2019.05.079

Girgis, H. Z. (2015). Red : An intelligent, rapid, accurate tool for detecting repeats de-novo on the genomic scale. *BMC Bioinformatics*, *16*(1), 227. https://doi.org/10.1186/s12859-015-0654-5

Golicz, A. A., Batley, J., & Edwards, D. (2016). Towards plant pangenomics. *Plant Biotechnology Journal*, *14*(4), 1099‐1105. https://doi.org/10.1111/pbi.12499

Golicz, A. A., Bayer, P. E., Bhalla, P. L., Batley, J., & Edwards, D. (2020). Pangenomics Comes of Age : From Bacteria to Plant and Animal Applications. *Trends in Genetics*, *36*(2), 132‐145. https://doi.org/10.1016/j.tig.2019.11.006

Golicz, A. A., Martinez, P. A., Zander, M., Patel, D. A., Van De Wouw, A. P., Visendi, P., Fitzgerald, T. L., Edwards, D., & Batley, J. (2015). Gene loss in the fungal canola pathogen Leptosphaeria maculans. *Functional & Integrative Genomics*, *15*(2), 189‐196. https://doi.org/10.1007/s10142-014-0412-1

Gordon, S. P., Contreras-Moreira, B., Woods, D. P., Des Marais, D. L., Burgess, D., Shu, S., Stritt, C., Roulin, A. C., Schackwitz, W., Tyler, L., Martin, J., Lipzen, A., Dochy, N., Phillips, J., Barry, K., Geuten, K., Budak, H., Juenger, T. E., Amasino, R., … Vogel, J. P. (2017). Extensive gene content variation in the Brachypodium distachyon pan-genome correlates with population structure. *Nature Communications*, *8*(1), 2184. https://doi.org/10.1038/s41467-017-02292-8

Gregory, T. R. (2009). Understanding Natural Selection : Essential Concepts and Common Misconceptions. *Evolution: Education and Outreach*, *2*(2), 156‐175. https://doi.org/10.1007/s12052-009-0128-1

Grigoriev, I. V., Nikitin, R., Haridas, S., Kuo, A., Ohm, R., Otillar, R., Riley, R., Salamov, A., Zhao, X., Korzeniewski, F., Smirnova, T., Nordberg, H., Dubchak, I., & Shabalov, I. (2014). MycoCosm portal : Gearing up for 1000 fungal genomes. *Nucleic Acids Research*, *42*(D1), D699‐D704. https://doi.org/10.1093/nar/gkt1183

Grimplet, J., Pimentel, D., Agudelo-Romero, P., Martinez-Zapater, J. M., & Fortes, A. M. (2017). The LATERAL ORGAN BOUNDARIES Domain gene family in grapevine : Genome-wide characterization and expression analyses during developmental processes and stress responses. *Scientific Reports*, *7*(1), 15968. https://doi.org/10.1038/s41598-017-16240-5

Grover, A., & Sharma, P. C. (2016). Development and use of molecular markers : Past and present. *Critical Reviews in Biotechnology*, *36*(2), 290‐302. https://doi.org/10.3109/07388551.2014.959891

Guo, L.-M., Li, J., He, J., Liu, H., & Zhang, H.-M. (2020). A class I cytosolic HSP20 of rice enhances heat and salt tolerance in different organisms. *Scientific Reports*, *10*(1), 1383. https://doi.org/10.1038/s41598-020-58395-8

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST : Quality assessment tool for genome assemblies. *Bioinformatics*, *29*(8), 1072-1075. https://doi.org/10.1093/bioinformatics/btt086

Habtemariam, S. (2019). Introduction to plant secondary metabolites—From biosynthesis to chemistry and antidiabetic action. In *Medicinal Foods as Potential Therapies for Type-2 Diabetes and Associated Diseases* (p. 109-132). Elsevier. https://doi.org/10.1016/B978-0-08-102922-0.00006-7

Hacquard, T., Clavel, M., Baldrich, P., Lechner, E., Pérez-Salamó, I., Schepetilnikov, M., Derrien, B., Dubois, M., Hammann, P., Kuhn, L., Brun, D., Bouteiller, N., Baumberger, N., Vaucheret, H., Meyers, B. C., & Genschik, P. (2022). The Arabidopsis F-box protein FBW2 targets AGO1 for degradation to prevent spurious loading of illegitimate small RNA. *Cell Reports*, *39*(2), 110671. https://doi.org/10.1016/j.celrep.2022.110671

Han, Z., Xiong, D., Schneiter, R., & Tian, C. (2023). The function of plant PR1 and other members of the CAP protein superfamily in plant–pathogen interactions. *Molecular Plant Pathology*, *24*(6), 651-668. https://doi.org/10.1111/mpp.13320

Hancock, A. M., Brachi, B., Faure, N., Horton, M. W., Jarymowycz, L. B., Sperone, F. G., Toomajian, C., Roux, F., & Bergelson, J. (2011). Adaptation to Climate Across the *Arabidopsis thaliana* Genome. *Science*, *334*(6052), 83-86. https://doi.org/10.1126/science.1209244

Harholt, J., Moestrup, Ø., & Ulvskov, P. (2016). Why Plants Were Terrestrial from the Beginning. *Trends in Plant Science*, *21*(2), 96-101. https://doi.org/10.1016/j.tplants.2015.11.010

Harris, B. J., Clark, J. W., Schrempf, D., Szöllősi, G. J., Donoghue, P. C. J., Hetherington, A. M., & Williams, T. A. (2022). Divergent evolutionary trajectories of bryophytes and tracheophytes from a complex common ancestor of land plants. *Nature Ecology & Evolution*, *6*(11), 1634-1643. https://doi.org/10.1038/s41559-022-01885-x

Harris, B. J., Harrison, C. J., Hetherington, A. M., & Williams, T. A. (2020). Phylogenomic Evidence for the Monophyly of Bryophytes and the Reductive Evolution of Stomata. *Current Biology*, *30*(11), 2001-2012.e2. https://doi.org/10.1016/j.cub.2020.03.048

Hartman, K., & Tringe, S. G. (2019). Interactions between plants and soil shaping the root microbiome under abiotic stress. *Biochemical Journal*, *476*(19), 2705-2724. https://doi.org/10.1042/BCJ20180615

He, Q., Tang, S., Zhi, H., Chen, J., Zhang, J., Liang, H., Alam, O., Li, H., Zhang, H., Xing, L., Li, X., Zhang, W., Wang, H., Shi, J., Du, H., Wu, H., Wang, L., Yang, P., Xing, L., … Diao, X. (2023). A graph-based genome and pan-genome variation of the model plant Setaria. *Nature Genetics*, *55*(7), 1232-1242. https://doi.org/10.1038/s41588-023-01423-w

Healey, A. L., Piatkowski, B., Lovell, J. T., Sreedasyam, A., Carey, S. B., Mamidi, S., Shu, S., Plott, C., Jenkins, J., Lawrence, T., Aguero, B., Carrell, A. A., Nieto-Lugilde, M., Talag, J., Duffy, A., Jawdy, S., Carter, K. R., Boston, L.-B., Jones, T., … Shaw, A. J. (2023). Newly identified sex chromosomes in the Sphagnum (peat moss) genome alter carbon sequestration and ecosystem dynamics. *Nature Plants*, *9*(2), 238-254. https://doi.org/10.1038/s41477-022-01333-5

Hedrick, P. W. (2007). Balancing selection. *Current Biology*, *17*(7), R230-R231. https://doi.org/10.1016/j.cub.2007.01.012

Hiwatashi, T., Goh, H., Yasui, Y., Koh, L. Q., Takami, H., Kajikawa, M., Kirita, H., Kanazawa, T., Minamino, N., Togawa, T., Sato, M., Wakazaki, M., Yamaguchi, K., Shigenobu, S., Fukaki, H., Mimura, T., Toyooka, K., Sawa, S., Yamato, K. T., … Ishizaki, K. (2019). The RopGEF KARAPPO Is Essential for the Initiation of

Vegetative Reproduction in Marchantia polymorpha. *Current Biology*, *29*(20), 3525-3531.e7. https://doi.org/10.1016/j.cub.2019.08.071

Hoey, D. J., Greiff, G. R. L., SLCU Outreach Consortium, & Schornack, S. (2023). *The Great British Liverwort Hunt – collecting wild accessions for molecular biology research while engaging the public* (v1.0) [jeu de données]. Zenodo. https://doi.org/10.5281/ZENODO.10040685

Hoff, K. J., Lange, S., Lomsadze, A., Borodovsky, M., & Stanke, M. (2016). BRAKER1 : Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics*, *32*(5), 767-769. https://doi.org/10.1093/bioinformatics/btv661

Hu, F., Ye, Z., Dong, K., Zhang, W., Fang, D., & Cao, J. (2023). Divergent structures and functions of the Cupin proteins in plants. *International Journal of Biological Macromolecules*, *242*, 124791. https://doi.org/10.1016/j.ijbiomac.2023.124791

Hu, H., Scheben, A., Verpaalen, B., Tirnaz, S., Bayer, P. E., Hodel, R. G. J., Batley, J., Soltis, D. E., Soltis, P. S., & Edwards, D. (2022). *Amborella* gene presence/absence variation is associated with abiotic stress responses that may contribute to environmental adaptation. *New Phytologist*, *233*(4), 1548-1555. https://doi.org/10.1111/nph.17658

Hu, S., Dilcher, D. L., Jarzen, D. M., & Winship Taylor, D. (2008). Early steps of angiosperm–pollinator coevolution. *Proceedings of the National Academy of Sciences*, *105*(1), 240-245. https://doi.org/10.1073/pnas.0707989105

Huang, J. (2013). Horizontal gene transfer in eukaryotes : The weak-link model. *BioEssays*, *35*(10), 868-875. https://doi.org/10.1002/bies.201300007

Hübner, S., Bercovich, N., Todesco, M., Mandel, J. R., Odenheimer, J., Ziegler, E., Lee, J. S., Baute, G. J., Owens, G. L., Grassa, C. J., Ebert, D. P., Ostevik, K. L., Moyers, B. T., Yakimowski, S., Masalia, R. R., Gao, L., Ćalić, I., Bowers, J. E., Kane, N. C., … Rieseberg, L. H. (2019). Sunflower pan-genome analysis shows that hybridization altered gene content and disease resistance. *Nature Plants*, *5*(1), 54-62. https://doi.org/10.1038/s41477-018-0329-0

Hudson, R. R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, *18*(2), 337-338. https://doi.org/10.1093/bioinformatics/18.2.337

Hufnagel, B., Soriano, A., Taylor, J., Divol, F., Kroc, M., Sanders, H., Yeheyis, L., Nelson, M., & Péret, B. (2021). Pangenome of white lupin provides insights into the diversity of the species. *Plant Biotechnology Journal*, *19*(12), 2532-2543. https://doi.org/10.1111/pbi.13678

Hyun, J. C., Monk, J. M., & Palsson, B. O. (2022). Comparative pangenomics : Analysis of 12 microbial pathogen pangenomes reveals conserved global structures of genetic and functional diversity. *BMC Genomics*, *23*(1), 7. https://doi.org/10.1186/s12864-021-08223-8

Intergovernmental Panel On Climate Change (Ipcc). (2023). *Climate Change 2022 – Impacts, Adaptation and Vulnerability : Working Group II Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* (1re éd.). Cambridge University Press. https://doi.org/10.1017/9781009325844

Ishizaki, K., Mizutani, M., Shimamura, M., Masuda, A., Nishihama, R., & Kohchi, T. (2013). Essential Role of the E3 Ubiquitin Ligase NOPPERABO1 in Schizogenous Intercellular Space Formation in the Liverwort *Marchantia polymorpha*. *The Plant Cell*, *25*(10), 4075-4084. https://doi.org/10.1105/tpc.113.117051

Ishizaki, K., Nishihama, R., Yamato, K. T., & Kohchi, T. (2016). Molecular Genetic Tools and Techniques for *Marchantia polymorpha* Research. *Plant and Cell Physiology*, *57*(2), 262‑270. https://doi.org/10.1093/pcp/pcv097

Jia, Q., Li, G., Köllner, T. G., Fu, J., Chen, X., Xiong, W., Crandall-Stotler, B. J., Bowman, J. L., Weston, D. J., Zhang, Y., Chen, L., Xie, Y., Li, F.-W., Rothfels, C. J., Larsson, A., Graham, S. W., Stevenson, D. W., Wong, G. K.-S., Gershenzon, J., & Chen, F. (2016). Microbial-type terpene synthase genes occur widely in nonseed land plants, but not in seed plants. *Proceedings of the National Academy of Sciences*, *113*(43), 12328‑12333. https://doi.org/10.1073/pnas.1607973113

Jombart, T., Devillard, S., & Balloux, F. (2010). Discriminant analysis of principal components : A new method for the analysis of genetically structured populations. *BMC Genetics*, *11*(1), 94. https://doi.org/10.1186/1471-2156-11-94

Joshi, R., Wani, S. H., Singh, B., Bohra, A., Dar, Z. A., Lone, A. A., Pareek, A., & Singla-Pareek, S. L. (2016). Transcription Factors and Plants Response to Drought Stress : Current Understanding and Future Directions. *Frontiers in Plant Science*, *7*. https://doi.org/10.3389/fpls.2016.01029

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., Von Haeseler, A., & Jermiin, L. S. (2017). ModelFinder : Fast model selection for accurate phylogenetic estimates. *Nature Methods*, *14*(6), 587‑589. https://doi.org/10.1038/nmeth.4285

Kanazawa, T., Morinaka, H., Ebine, K., Shimada, T. L., Ishida, S., Minamino, N., Yamaguchi, K., Shigenobu, S., Kohchi, T., Nakano, A., & Ueda, T. (2020). The liverwort oil body is formed by redirection of the secretory pathway. *Nature Communications*, *11*(1), 6152. https://doi.org/10.1038/s41467-020-19978-1

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S., Freimer, N. B., Sabatti, C., & Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, *42*(4), 348‑354. https://doi.org/10.1038/ng.548

Karasov, T. L., Chae, E., Herman, J. J., & Bergelson, J. (2017). Mechanisms to Mitigate the Trade-Off between Growth and Defense. *The Plant Cell*, *29*(4), 666‑680. https://doi.org/10.1105/tpc.16.00931

Karasov, T. L., Horton, M. W., & Bergelson, J. (2014). Genomic variability as a driver of plant–pathogen coevolution? *Current Opinion in Plant Biology*, *18*, 24‑30. https://doi.org/10.1016/j.pbi.2013.12.003

Kato, H., Yasui, Y., & Ishizaki, K. (2020). Gemma cup and gemma development in *Marchantia polymorpha*. *New Phytologist*, *228*(2), 459‑465. https://doi.org/10.1111/nph.16655

Khatun, N., Shinozawa, A., Takahashi, K., Matsuura, H., Jahan, A., Islam, M., Karim, M., Sk, R., Yoshikawa, M., Ishizaki, K., Sakata, Y., & Takezawa, D. (2023). Abscisic acid-mediated sugar responses are essential for vegetative desiccation tolerance in the liverwort *Marchantia polymorpha*. *Physiologia Plantarum*, *175*(2), e13898. https://doi.org/10.1111/ppl.13898

Kidwai, M., Ahmad, I. Z., & Chakrabarty, D. (2020). Class III peroxidase : An indispensable enzyme for biotic/abiotic stress tolerance and a potent candidate for crop improvement. *Plant Cell Reports*, *39*(11), 1381‑1393. https://doi.org/10.1007/s00299-020-02588-y

Kimura, M. (1991). The neutral theory of molecular evolution : A review of recent evidence. *The Japanese Journal of Genetics*, *66*(4), 367‑386. https://doi.org/10.1266/jjg.66.367

Kingman, J. F. C. (2000). Origins of the Coalescent : 1974-1982. *Genetics*, *156*(4), 1461‑1463. https://doi.org/10.1093/genetics/156.4.1461

Koboldt, D. C., Chen, K., Wylie, T., Larson, D. E., McLellan, M. D., Mardis, E. R., Weinstock, G. M., Wilson, R. K., & Ding, L. (2009). VarScan : Variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics*, *25*(17), 2283‑2285. https://doi.org/10.1093/bioinformatics/btp373

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., Miller, C. A., Mardis, E. R., Ding, L., & Wilson, R. K. (2012). VarScan 2 : Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*, *22*(3), 568‑576. https://doi.org/10.1101/gr.129684.111

Kohchi, T., Yamato, K. T., Ishizaki, K., Yamaoka, S., & Nishihama, R. (2021). Development and Molecular Genetics of *Marchantia polymorpha*. *Annual Review of Plant Biology*, *72*(1), 677‑702. https://doi.org/10.1146/annurev-arplant-082520-094256

Kolkas, H., Balliau, T., Chourré, J., Zivy, M., Canut, H., & Jamet, E. (2022). The Cell Wall Proteome of Marchantia polymorpha Reveals Specificities Compared to Those of Flowering Plants. *Frontiers in Plant Science*, *12*, 765846. https://doi.org/10.3389/fpls.2021.765846

Kong, L., Liu, Y., Zhi, P., Wang, X., Xu, B., Gong, Z., & Chang, C. (2020). Origins and Evolution of Cuticle Biosynthetic Machinery in Land Plants. *Plant Physiology*, *184*(4), 1998‑2010. https://doi.org/10.1104/pp.20.00913

Kremer, A., & Hipp, A. L. (2020). Oaks : An evolutionary success story. *New Phytologist*, *226*(4), 987‑1011. https://doi.org/10.1111/nph.16274

Krueger, F., James, F., Ewels, P., Afyounian, E., & Schuster-Boeckler, B. (2021). *FelixKrueger/TrimGalore : V0.6.7 - DOI via Zenodo* (0.6.7) [Logiciel]. Zenodo. https://doi.org/10.5281/ZENODO.5127899

Kumar, S., Kempinski, C., Zhuang, X., Norris, A., Mafu, S., Zi, J., Bell, S. A., Nybo, S. E., Kinison, S. E., Jiang, Z., Goklany, S., Linscott, K. B., Chen, X., Jia, Q., Brown, S. D., Bowman, J. L., Babbitt, P. C., Peters, R. J., Chen, F., & Chappell, J. (2016). Molecular Diversity of Terpene Synthases in the Liverwort Marchantia polymorpha. *The Plant Cell*, tpc.00062.2016. https://doi.org/10.1105/tpc.16.00062

Kunnev, D. (2020). Origin of Life : The Point of No Return. *Life*, *10*(11), 269. https://doi.org/10.3390/life10110269

Lang, D., Ullrich, K. K., Murat, F., Fuchs, J., Jenkins, J., Haas, F. B., Piednoel, M., Gundlach, H., Van Bel, M., Meyberg, R., Vives, C., Morata, J., Symeonidi, A., Hiss, M., Muchero, W., Kamisugi, Y., Saleh, O., Blanc, G., Decker, E. L., … Rensing, S. A. (2018). The *Physcomitrella patens* chromosome-scale assembly reveals moss genome structure and evolution. *The Plant Journal*, *93*(3), 515‑533. https://doi.org/10.1111/tpj.13801

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357‑359. https://doi.org/10.1038/nmeth.1923

Lanot, A., Hodge, D., Jackson, R. G., George, G. L., Elias, L., Lim, E., Vaistij, F. E., & Bowles, D. J. (2006). The glucosyltransferase UGT72E2 is responsible for monolignol 4- *O* -glucoside production in *Arabidopsis thaliana*. *The Plant Journal*, *48*(2), 286‑295. https://doi.org/10.1111/j.1365-313X.2006.02872.x

Lasky, J. R., Josephs, E. B., & Morris, G. P. (2023). Genotype–environment associations to reveal the molecular basis of environmental adaptation. *The Plant Cell*, *35*(1), 125-138. https://doi.org/10.1093/plcell/koac267

Lenton, T. M., Dahl, T. W., Daines, S. J., Mills, B. J. W., Ozaki, K., Saltzman, M. R., & Porada, P. (2016). Earliest land plants created modern levels of atmospheric oxygen. *Proceedings of the National Academy of Sciences*, *113*(35), 9704-9709. https://doi.org/10.1073/pnas.1604787113

Letunic, I., & Bork, P. (2021). Interactive Tree Of Life (iTOL) v5 : An online tool for phylogenetic tree display and annotation. *Nucleic Acids Research*, *49*(W1), W293-W296. https://doi.org/10.1093/nar/gkab301

Li, D., Liu, C.-M., Luo, R., Sadakane, K., & Lam, T.-W. (2015). MEGAHIT : An ultra-fast single-node solution for large and complex metagenomics assembly via succinct *de Bruijn* graph. *Bioinformatics*, *31*(10), 1674-1676. https://doi.org/10.1093/bioinformatics/btv033

Li, F.-W., Nishiyama, T., Waller, M., Frangedakis, E., Keller, J., Li, Z., Fernandez-Pozo, N., Barker, M. S., Bennett, T., Blázquez, M. A., Cheng, S., Cuming, A. C., de Vries, J., de Vries, S., Delaux, P.-M., Diop, I. S., Harrison, C. J., Hauser, D., Hernández-García, J., … Szövényi, P. (2020). Anthoceros genomes illuminate the origin of land plants and the unique biology of hornworts. *Nature Plants*, *6*(3), 259-272. https://doi.org/10.1038/s41477-020-0618-2

Li, H., Feng, X., & Chu, C. (2020). The design and construction of reference pangenome graphs with minigraph. *Genome Biology*, *21*(1), 265. https://doi.org/10.1186/s13059-020-02168-z

Li, H., Wang, S., Chai, S., Yang, Z., Zhang, Q., Xin, H., Xu, Y., Lin, S., Chen, X., Yao, Z., Yang, Q., Fei, Z., Huang, S., & Zhang, Z. (2022). Graph-based pan-genome reveals structural and sequence variations related to agronomic traits and domestication in cucumber. *Nature Communications*, *13*(1), 682. https://doi.org/10.1038/s41467-022-28362-0

Li, J., Yuan, D., Wang, P., Wang, Q., Sun, M., Liu, Z., Si, H., Xu, Z., Ma, Y., Zhang, B., Pei, L., Tu, L., Zhu, L., Chen, L.-L., Lindsey, K., Zhang, X., Jin, S., & Wang, M. (2021). Cotton pan-genome retrieves the lost sequences and genes during domestication and selection. *Genome Biology*, *22*(1), 119. https://doi.org/10.1186/s13059-021-02351-w

Li, P., Quan, X., Jia, G., Xiao, J., Cloutier, S., & You, F. M. (2016). RGAugury : A pipeline for genome-wide prediction of resistance gene analogs (RGAs) in plants. *BMC Genomics*, *17*(1), 852. https://doi.org/10.1186/s12864-016-3197-x

Li, S., Liu, S., Zhang, Q., Cui, M., Zhao, M., Li, N., Wang, S., Wu, R., Zhang, L., Cao, Y., & Wang, L. (2022). The interaction of ABA and ROS in plant growth and stress resistances. *Frontiers in Plant Science*, *13*, 1050132. https://doi.org/10.3389/fpls.2022.1050132

Li, W., & Godzik, A. (2006). Cd-hit : A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, *22*(13), 1658-1659. https://doi.org/10.1093/bioinformatics/btl158

Li, Y., Zhou, G., Ma, J., Jiang, W., Jin, L., Zhang, Z., Guo, Y., Zhang, J., Sui, Y., Zheng, L., Zhang, S., Zuo, Q., Shi, X., Li, Y., Zhang, W., Hu, Y., Kong, G., Hong, H., Tan, B., … Qiu, L. (2014). De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nature Biotechnology*, *32*(10), 1045-1052. https://doi.org/10.1038/nbt.2979

Linde, A.-M., Eklund, D. M., Cronberg, N., Bowman, J. L., & Lagercrantz, U. (2021). Rates and patterns of molecular evolution in bryophyte genomes, with focus on complex thalloid liverworts, Marchantiopsida. *Molecular Phylogenetics and Evolution*, *165*, 107295. https://doi.org/10.1016/j.ympev.2021.107295

Linde, A.-M., Sawangproh, W., Cronberg, N., Szövényi, P., & Lagercrantz, U. (2020). Evolutionary History of the Marchantia polymorpha Complex. *Frontiers in Plant Science*, *11*, 829. https://doi.org/10.3389/fpls.2020.00829

Linde, A.-M., Singh, S., Bowman, J. L., Eklund, M., Cronberg, N., & Lagercrantz, U. (2023). Genome Evolution in Plants : Complex Thalloid Liverworts (Marchantiopsida). *Genome Biology and Evolution*, *15*(3), evad014. https://doi.org/10.1093/gbe/evad014

Linkies, A., Graeber, K., Knight, C., & Leubner-Metzger, G. (2010). The evolution of seeds. *New Phytologist*, *186*(4), 817-831. https://doi.org/10.1111/j.1469-8137.2010.03249.x

Liu, C., Wang, Y., Peng, J., Fan, B., Xu, D., Wu, J., Cao, Z., Gao, Y., Wang, X., Li, S., Su, Q., Zhang, Z., Wang, S., Wu, X., Shang, Q., Shi, H., Shen, Y., Wang, B., & Tian, J. (2022). High-quality genome assembly and pan-genome studies facilitate genetic discovery in mung bean and its improvement. *Plant Communications*, *3*(6), 100352. https://doi.org/10.1016/j.xplc.2022.100352

Liu, Y., Du, H., Li, P., Shen, Y., Peng, H., Liu, S., Zhou, G.-A., Zhang, H., Liu, Z., Shi, M., Huang, X., Li, Y., Zhang, M., Wang, Z., Zhu, B., Han, B., Liang, C., & Tian, Z. (2020). Pan-Genome of Wild and Cultivated Soybeans. *Cell*, *182*(1), 162-176.e13. https://doi.org/10.1016/j.cell.2020.05.023

Liu, Y., Zeng, Z., Zhang, Y.-M., Li, Q., Jiang, X.-M., Jiang, Z., Tang, J.-H., Chen, D., Wang, Q., Chen, J.-Q., & Shao, Z.-Q. (2021). An angiosperm NLR Atlas reveals that NLR gene reduction is associated with ecological specialization and signal transduction component deletion. *Molecular Plant*, *14*(12), 2015-2031. https://doi.org/10.1016/j.molp.2021.08.001

Lofgren, L. A., Ross, B. S., Cramer, R. A., & Stajich, J. E. (2022). The pan-genome of Aspergillus fumigatus provides a high-resolution view of its population structure revealing high levels of lineage-specific diversity driven by recombination. *PLOS Biology*, *20*(11), e3001890. https://doi.org/10.1371/journal.pbio.3001890

Lundquist, P. K., Davis, J. I., & Van Wijk, K. J. (2012). ABC1K atypical kinases in plants : Filling the organellar kinase void. *Trends in Plant Science*, *17*(9), 546-555. https://doi.org/10.1016/j.tplants.2012.05.010

Ma, J., Wang, S., Zhu, X., Sun, G., Chang, G., Li, L., Hu, X., Zhang, S., Zhou, Y., Song, C.-P., & Huang, J. (2022). Major episodes of horizontal gene transfer drove the evolution of land plants. *Molecular Plant*, *15*(5), 857-871. https://doi.org/10.1016/j.molp.2022.02.001

Malivert, A., & Hamant, O. (2023). Why is FERONIA pleiotropic? *Nature Plants*, *9*(7), 1018-1025. https://doi.org/10.1038/s41477-023-01434-9

Manara, A., DalCorso, G., & Furini, A. (2016). The Role of the Atypical Kinases ABC1K7 and ABC1K8 in Abscisic Acid Responses. *Frontiers in Plant Science*, *7*. https://doi.org/10.3389/fpls.2016.00366

Manara, A., DalCorso, G., Guzzo, F., & Furini, A. (2015). Loss of the Atypical Kinases ABC1K7 and ABC1K8 Changes the Lipid Composition of the Chloroplast Membrane. *Plant and Cell Physiology*, *56*(6), 1193-1204. https://doi.org/10.1093/pcp/pcv046

Manara, A., DalCorso, G., Leister, D., Jahns, P., Baldan, B., & Furini, A. (2014). At SIA 1 AND At OSA 1 : Two Abc1 proteins involved in oxidative stress responses and iron distribution within chloroplasts. *New Phytologist*, *201*(2), 452-465. https://doi.org/10.1111/nph.12533

Manni, M., Berkeley, M. R., Seppey, M., & Zdobnov, E. M. (2021). BUSCO : Assessing Genomic Data Quality and Beyond. *Current Protocols*, *1*(12), e323. https://doi.org/10.1002/cpz1.323

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., … Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, *461*(7265), 747-753. https://doi.org/10.1038/nature08494

Manzano, D., Andrade, P., Caudepón, D., Altabella, T., Arró, M., & Ferrer, A. (2016). Suppressing Farnesyl Diphosphate Synthase Alters Chloroplast Development and Triggers Sterol-Dependent Induction of Jasmonate- and Fe-Related Responses. *Plant Physiology*, *172*(1), 93-117. https://doi.org/10.1104/pp.16.00431

Mapleson, D., Garcia Accinelli, G., Kettleborough, G., Wright, J., & Clavijo, B. J. (2017). KAT : A K-mer analysis toolkit to quality control NGS datasets and genome assemblies. *Bioinformatics*, *33*(4), 574-576. https://doi.org/10.1093/bioinformatics/btw663

Marble, S. C., Frank, M. S., Laughinghouse, D. D., Steed, S. T., & Boyd, N. S. (2017). Biology and Management of Liverwort (Marchantia polymorpha) in Ornamental Crop Production. *EDIS*, *2017*(5). https://doi.org/10.32473/edis-ep542-2017

Mariette, J., Blanchard, O., Berné, O., Aumont, O., Carrey, J., Ligozat, A., Lellouch, E., Roche, P.-E., Guennebaud, G., Thanwerdas, J., Bardou, P., Salin, G., Maigne, E., Servan, S., & Ben-Ari, T. (2022). An open-source tool to assess the carbon footprint of research. *Environmental Research: Infrastructure and Sustainability*, *2*(3), 035008. https://doi.org/10.1088/2634-4505/ac84a4

Marks, R. A., Amézquita, E. J., Percival, S., Rougon-Cardoso, A., Chibici-Revneanu, C., Tebele, S. M., Farrant, J. M., Chitwood, D. H., & VanBuren, R. (2023). A critical analysis of plant science literature reveals ongoing inequities. *Proceedings of the National Academy of Sciences*, *120*(10), e2217564120. https://doi.org/10.1073/pnas.2217564120

Marks, R. A., Hotaling, S., Frandsen, P. B., & VanBuren, R. (2021). Representation and participation across 20 years of plant genome sequencing. *Nature Plants*, *7*(12), 1571-1578. https://doi.org/10.1038/s41477-021-01031-8

Marks, R. A., Smith, J. J., Cronk, Q., Grassa, C. J., & McLetchie, D. N. (2019). Genome of the tropical plant Marchantia inflexa : Implications for sex chromosome evolution and dehydration tolerance. *Scientific Reports*, *9*(1), 8722. https://doi.org/10.1038/s41598-019-45039-9

Marowa, P., Ding, A., & Kong, Y. (2016). Expansins : Roles in plant growth and potential applications in crop improvement. *Plant Cell Reports*, *35*(5), 949-965. https://doi.org/10.1007/s00299-016-1948-4

Masle, J., Gilmore, S. R., & Farquhar, G. D. (2005). The ERECTA gene regulates plant transpiration efficiency in Arabidopsis. *Nature*, *436*(7052), 866-870. https://doi.org/10.1038/nature03835

Mata-Pérez, C., & Spoel, S. H. (2019). Thioredoxin-mediated redox signalling in plant immunity. *Plant Science*, *279*, 27-33. https://doi.org/10.1016/j.plantsci.2018.05.001

Mateo-Bonmatí, E., Esteve-Bruna, D., Juan-Vicente, L., Nadi, R., Candela, H., Lozano, F. M., Ponce, M. R., Pérez-Pérez, J. M., & Micol, J. L. (2018). *INCURVATA11* and *CUPULIFORMIS2* Are Redundant Genes

That Encode Epigenetic Machinery Components in Arabidopsis. *The Plant Cell*, *30*(7), 1596-1616. https://doi.org/10.1105/tpc.18.00300

Matsui, H., Iwakawa, H., Hyon, G.-S., Yotsui, I., Katou, S., Monte, I., Nishihama, R., Franzen, R., Solano, R., & Nakagami, H. (2020). Isolation of Natural Fungal Pathogens from Marchantia polymorpha Reveals Antagonism between Salicylic Acid and Jasmonate during Liverwort–Fungus Interactions. *Plant and Cell Physiology*, *61*(2), 265-275. https://doi.org/10.1093/pcp/pcz187

Maurel, C., Boursiac, Y., Luu, D.-T., Santoni, V., Shahzad, Z., & Verdoucq, L. (2015). Aquaporins in Plants. *Physiological Reviews*, *95*(4), 1321-1358. https://doi.org/10.1152/physrev.00008.2015

McDaniel, S. F. (2021). Bryophytes are not early diverging land plants. *New Phytologist*, nph.17241. https://doi.org/10.1111/nph.17241

Mecchia, M. A., Rövekamp, M., Giraldo-Fonseca, A., Meier, D., Gadient, P., Vogler, H., Limacher, D., Bowman, J. L., & Grossniklaus, U. (2022). The single *Marchantia polymorpha FERONIA* homolog reveals an ancestral role in regulating cellular expansion and integrity. *Development*, *149*(19), dev200580. https://doi.org/10.1242/dev.200580

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2 : New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, *37*(5), 1530-1534. https://doi.org/10.1093/molbev/msaa015

Monaghan, F., & Corcos, A. (1984). On the origins of the Mendelian laws. *Journal of Heredity*, *75*(1), 67-69. https://doi.org/10.1093/oxfordjournals.jhered.a109868

Monroe, J. G., McKay, J. K., Weigel, D., & Flood, P. J. (2021). The population genomics of adaptive loss of function. *Heredity*, *126*(3), 383-395. https://doi.org/10.1038/s41437-021-00403-2

Monte, I. (2023). Jasmonates and salicylic acid : Evolution of defense hormones in land plants. *Current Opinion in Plant Biology*, *76*, 102470. https://doi.org/10.1016/j.pbi.2023.102470

Montgomery, S. A., & Berger, F. (2023). Paternal imprinting in *Marchantia polymorpha*. *New Phytologist*, nph.19377. https://doi.org/10.1111/nph.19377

Montgomery, S. A., Tanizawa, Y., Galik, B., Wang, N., Ito, T., Mochizuki, T., Akimcheva, S., Bowman, J. L., Cognat, V., Maréchal-Drouard, L., Ekker, H., Hong, S.-F., Kohchi, T., Lin, S.-S., Liu, L.-Y. D., Nakamura, Y., Valeeva, L. R., Shakirov, E. V., Shippen, D. E., … Berger, F. (2020). Chromatin Organization in Early Land Plants Reveals an Ancestral Association between H3K27me3, Transposons, and Constitutive Heterochromatin. *Current Biology*, *30*(4), 573-588.e7. https://doi.org/10.1016/j.cub.2019.12.015

Morgante, M., Depaoli, E., & Radovic, S. (2007). Transposable elements and the plant pan-genomes. *Current Opinion in Plant Biology*, *10*(2), 149-155. https://doi.org/10.1016/j.pbi.2007.02.001

Nelsen, M. P., Lücking, R., Boyce, C. K., Lumbsch, H. T., & Ree, R. H. (2020). No support for the emergence of lichens prior to the evolution of vascular plants. *Geobiology*, *18*(1), 3-13. https://doi.org/10.1111/gbi.12369

Nelson, J. M., Hauser, D. A., Hinson, R., & Shaw, A. J. (2018). A novel experimental system using the liverwort *Marchantia polymorpha* and its fungal endophytes reveals diverse and context-dependent effects. *New Phytologist*, *218*(3), 1217-1232. https://doi.org/10.1111/nph.15012

Nelson, J., & Shaw, A. J. (2019). Exploring the natural microbiome of the model liverwort : Fungal endophyte diversity in Marchantia polymorpha L. *Symbiosis*, *78*(1), 45-59. https://doi.org/10.1007/s13199-019-00597-4

Nielsen, R. (1997). A Likelihood Approach to Populations Samples of Microsatellite Alleles. *Genetics*, *146*(2), 711-716. https://doi.org/10.1093/genetics/146.2.711

Nishimura, K., Apitz, J., Friso, G., Kim, J., Ponnala, L., Grimm, B., & Van Wijk, K. J. (2015). Discovery of a Unique Clp Component, ClpF, in Chloroplasts : A Proposed Binary ClpF-ClpS1 Adaptor Complex Functions in Substrate Recognition and Delivery. *The Plant Cell*, tpc.15.00574. https://doi.org/10.1105/tpc.15.00574

Ohno, S. (1970). *Evolution by Gene Duplication*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-86659-3

Ohyama, K., Fukuzawa, H., Kohchi, T., Shirai, H., Sano, T., Sano, S., Umesono, K., Shiki, Y., Takeuchi, M., Chang, Z., Aota, S., Inokuchi, H., & Ozeki, H. (1986). Chloroplast gene organization deduced from complete sequence of liverwort Marchantia polymorpha chloroplast DNA. *Nature*, *322*(6079), 572-574. https://doi.org/10.1038/322572a0

Oldroyd, D. R. (1986). Charles Darwin's theory of evolution : A review of our present understanding. *Biology & Philosophy*, *1*(2), 133-168. https://doi.org/10.1007/BF00142899

Olson, M. V. (1999). When Less Is More : Gene Loss as an Engine of Evolutionary Change. *The American Journal of Human Genetics*, *64*(1), 18-23. https://doi.org/10.1086/302219

O'Malley, M. A., Wideman, J. G., & Ruiz-Trillo, I. (2016). Losing Complexity : The Role of Simplification in Macroevolution. *Trends in Ecology & Evolution*, *31*(8), 608-621. https://doi.org/10.1016/j.tree.2016.04.004

One Thousand Plant Transcriptomes Initiative. (2019). One thousand plant transcriptomes and the phylogenomics of green plants. *Nature*, *574*(7780), 679-685. https://doi.org/10.1038/s41586-019-1693-2

Otto, S. P., & Goldstein, D. B. (1992). Recombination and the evolution of diploidy. *Genetics*, *131*(3), 745-751. https://doi.org/10.1093/genetics/131.3.745

Ozaki, K., Ohnishi, Y., Iida, A., Sekine, A., Yamada, R., Tsunoda, T., Sato, H., Sato, H., Hori, M., Nakamura, Y., & Tanaka, T. (2002). Functional SNPs in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction. *Nature Genetics*, *32*(4), 650-654. https://doi.org/10.1038/ng1047

Parks, M. M., Lawrence, C. E., & Raphael, B. J. (2015). Detecting non-allelic homologous recombination from high-throughput sequencing data. *Genome Biology*, *16*(1), 72. https://doi.org/10.1186/s13059-015-0633-1

Passardi, F., Theiler, G., Zamocky, M., Cosio, C., Rouhier, N., Teixera, F., Margis-Pinheiro, M., Ioannidis, V., Penel, C., Falquet, L., & Dunand, C. (2007). PeroxiBase : The peroxidase database. *Phytochemistry*, *68*(12), 1605-1611. https://doi.org/10.1016/j.phytochem.2007.04.005

Pecrix, Y., Staton, S. E., Sallet, E., Lelandais-Brière, C., Moreau, S., Carrère, S., Blein, T., Jardinaud, M.-F., Latrasse, D., Zouine, M., Zahm, M., Kreplak, J., Mayjonade, B., Satgé, C., Perez, M., Cauet, S., Marande, W., Chantry-Darmon, C., Lopez-Roques, C., … Gamas, P. (2018). Whole-genome landscape of

Medicago truncatula symbiotic genes. *Nature Plants*, *4*(12), 1017-1025. https://doi.org/10.1038/s41477-018-0286-7

Pederson, E. R. A., Warshan, D., & Rasmussen, U. (2019). Genome Sequencing of *Pleurozium schreberi* : The Assembled and Annotated Draft Genome of a Pleurocarpous Feather Moss. *G3 Genes|Genomes|Genetics*, *9*(9), 2791-2797. https://doi.org/10.1534/g3.119.400279

Peumans, W. J., Fouquaert, E., Jauneau, A., Rougé, P., Lannoo, N., Hamada, H., Alvarez, R., Devreese, B., & Van Damme, E. J. M. (2007). The Liverwort *Marchantia polymorpha* Expresses Orthologs of the Fungal *Agaricus bisporus* Agglutinin Family. *Plant Physiology*, *144*(2), 637-647. https://doi.org/10.1104/pp.106.087437

Porada, P., Lenton, T. M., Pohl, A., Weber, B., Mander, L., Donnadieu, Y., Beer, C., Pöschl, U., & Kleidon, A. (2016). High potential for weathering and climate effects of non-vascular vegetation in the Late Ordovician. *Nature Communications*, *7*(1), 12113. https://doi.org/10.1038/ncomms12113

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of Population Structure Using Multilocus Genotype Data. *Genetics*, *155*(2), 945.

Proctor, M. C. F., & Tuba, Z. (2002). Poikilohydry and homoihydry : Antithesis or spectrum of possibilities? *New Phytologist*, *156*(3), 327-349. https://doi.org/10.1046/j.1469-8137.2002.00526.x

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., De Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK : A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, *81*(3), 559-575. https://doi.org/10.1086/519795

Puttick, M. N., Morris, J. L., Williams, T. A., Cox, C. J., Edwards, D., Kenrick, P., Pressel, S., Wellman, C. H., Schneider, H., Pisani, D., & Donoghue, P. C. J. (2018). The Interrelationships of Land Plants and the Nature of the Ancestral Embryophyte. *Current Biology*, *28*(5), 733-745.e2. https://doi.org/10.1016/j.cub.2018.01.063

Radhakrishnan, G. V., Keller, J., Rich, M. K., Vernié, T., Mbadinga Mbadinga, D. L., Vigneron, N., Cottret, L., Clemente, H. S., Libourel, C., Cheema, J., Linde, A.-M., Eklund, D. M., Cheng, S., Wong, G. K. S., Lagercrantz, U., Li, F.-W., Oldroyd, G. E. D., & Delaux, P.-M. (2020). An ancestral signalling pathway is conserved in intracellular symbioses-forming plant lineages. *Nature Plants*, *6*(3), 280-289. https://doi.org/10.1038/s41477-020-0613-7

Raj, A., Stephens, M., & Pritchard, J. K. (2014). fastSTRUCTURE : Variational Inference of Population Structure in Large SNP Data Sets. *Genetics*, *197*(2), 573-589. https://doi.org/10.1534/genetics.114.164350

Redkar, A., Gimenez Ibanez, S., Sabale, M., Zechmann, B., Solano, R., & Di Pietro, A. (2022). *Marchantia polymorpha* model reveals conserved infection mechanisms in the vascular wilt fungal pathogen *Fusarium oxysporum*. *New Phytologist*, *234*(1), 227-241. https://doi.org/10.1111/nph.17909

Renault, H., Alber, A., Horst, N. A., Basilio Lopes, A., Fich, E. A., Kriegshauser, L., Wiedemann, G., Ullmann, P., Herrgott, L., Erhardt, M., Pineau, E., Ehlting, J., Schmitt, M., Rose, J. K. C., Reski, R., & Werck-Reichhart, D. (2017). A phenol-enriched cuticle is ancestral to lignin evolution in land plants. *Nature Communications*, *8*(1), 14713. https://doi.org/10.1038/ncomms14713

Rensing, S. A. (2018). Great moments in evolution : The conquest of land by plants. *Current Opinion in Plant Biology*, *42*, 49-54. https://doi.org/10.1016/j.pbi.2018.02.006

Rensing, S. A., Lang, D., Zimmer, A. D., Terry, A., Salamov, A., Shapiro, H., Nishiyama, T., Perroud, P.-F., Lindquist, E. A., Kamisugi, Y., Tanahashi, T., Sakakibara, K., Fujita, T., Oishi, K., Shin-I, T., Kuroki, Y., Toyoda, A., Suzuki, Y., Hashimoto, S. -i., … Boore, J. L. (2008). The Physcomitrella Genome Reveals Evolutionary Insights into the Conquest of Land by Plants. *Science*, *319*(5859), 64-69. https://doi.org/10.1126/science.1150646

Rey, T., Bonhomme, M., Chatterjee, A., Gavrin, A., Toulotte, J., Yang, W., André, O., Jacquet, C., & Schornack, S. (2017). The Medicago truncatula GRAS protein RAD1 supports arbuscular mycorrhiza symbiosis and Phytophthora palmivora susceptibility. *Journal of Experimental Botany*, *68*(21-22), 5871-5881. https://doi.org/10.1093/jxb/erx398

Rich, M. K., & Delaux, P.-M. (2020). Plant Evolution : When Arabidopsis Is More Ancestral Than Marchantia. *Current Biology*, *30*(11), R642-R644. https://doi.org/10.1016/j.cub.2020.04.077

Rich, M. K., Vigneron, N., Libourel, C., Keller, J., Xue, L., Hajheidari, M., Radhakrishnan, G. V., Le Ru, A., Diop, S. I., Potente, G., Conti, E., Duijsings, D., Batut, A., Le Faouder, P., Kodama, K., Kyozuka, J., Sallet, E., Bécard, G., Rodriguez-Franco, M., … Delaux, P.-M. (2021). Lipid exchanges drove the evolution of mutualism during plant terrestrialization. *Science*, *372*(6544), 864-868. https://doi.org/10.1126/science.abg0929

Romani, F., Banić, E., Florent, S. N., Kanazawa, T., Goodger, J. Q. D., Mentink, R. A., Dierschke, T., Zachgo, S., Ueda, T., Bowman, J. L., Tsiantis, M., & Moreno, J. E. (2020). Oil Body Formation in Marchantia polymorpha Is Controlled by MpC1HDZ and Serves as a Defense against Arthropod Herbivores. *Current Biology*, *30*(14), 2815-2828.e8. https://doi.org/10.1016/j.cub.2020.05.081

Romani, F., Flores, J. R., Tolopka, J. I., Suárez, G., He, X., & Moreno, J. E. (2022). Liverwort oil bodies : Diversity, biochemistry, and molecular cell biology of the earliest secretory structure of land plants. *Journal of Experimental Botany*, *73*(13), 4427-4439. https://doi.org/10.1093/jxb/erac134

Saijo, Y., & Loo, E. P. (2020). Plant immunity in signal integration between biotic and abiotic stress responses. *New Phytologist*, *225*(1), 87-104. https://doi.org/10.1111/nph.15989

Salamov, A. A., & Solovyev, V. V. (2000). Ab initio Gene Finding in *Drosophila* Genomic DNA. *Genome Research*, *10*(4), 516-522. https://doi.org/10.1101/gr.10.4.516

Sandler, G., Agrawal, A. F., & Wright, S. I. (2023). Population genomics of the facultatively sexual liverwort *Marchantia polymorpha*. *Genome Biology and Evolution*, evad196. https://doi.org/10.1093/gbe/evad196

Savelli, B., Li, Q., Webber, M., Jemmat, A. M., Robitaille, A., Zamocky, M., Mathé, C., & Dunand, C. (2019). RedoxiBase : A database for ROS homeostasis regulated proteins. *Redox Biology*, *26*, 101247. https://doi.org/10.1016/j.redox.2019.101247

Saxena, R. K., Edwards, D., & Varshney, R. K. (2014). Structural variations in plant genomes. *Briefings in Functional Genomics*, *13*(4), 296-307. https://doi.org/10.1093/bfgp/elu016

Schmid-Siegert, E., Sarkar, N., Iseli, C., Calderon, S., Gouhier-Darimont, C., Chrast, J., Cattaneo, P., Schütz, F., Farinelli, L., Pagni, M., Schneider, M., Voumard, J., Jaboyedoff, M., Fankhauser, C., Hardtke, C. S., Keller, L., Pannell, J. R., Reymond, A., Robinson-Rechavi, M., … Reymond, P. (2017). Low number of fixed somatic mutations in a long-lived oak tree. *Nature Plants*, *3*(12), 926-929. https://doi.org/10.1038/s41477-017-0066-9

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003). Cytoscape : A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, *13*(11), 2498-2504. https://doi.org/10.1101/gr.1239303

Shimamura, M. (2016). *Marchantia polymorpha* : Taxonomy, Phylogeny and Morphology of a Model System. *Plant and Cell Physiology*, *57*(2), 230-256. https://doi.org/10.1093/pcp/pcv192

Shortlidge, E. E., Carey, S. B., Payton, A. C., McDaniel, S. F., Rosenstiel, T. N., & Eppley, S. M. (2021). Microarthropod contributions to fitness variation in the common moss *Ceratodon purpureus*. *Proceedings of the Royal Society B: Biological Sciences*, *288*(1947), rspb.2021.0119, 20210119. https://doi.org/10.1098/rspb.2021.0119

Singh, B. K., Delgado-Baquerizo, M., Egidi, E., Guirado, E., Leach, J. E., Liu, H., & Trivedi, P. (2023). Climate change impacts on plant pathogens, food security and paths forward. *Nature Reviews Microbiology*, *21*(10), 640-656. https://doi.org/10.1038/s41579-023-00900-7

Singh, B., & Sharma, R. A. (2015). Plant terpenes : Defense responses, phylogenetic analysis, regulation and clinical applications. *3 Biotech*, *5*(2), 129-151. https://doi.org/10.1007/s13205-014-0220-2

Singh, M. K., Krüger, F., Beckmann, H., Brumm, S., Vermeer, J. E. M., Munnik, T., Mayer, U., Stierhof, Y.-D., Grefen, C., Schumacher, K., & Jürgens, G. (2014). Protein Delivery to Vacuole Requires SAND Protein-Dependent Rab GTPase Conversion for MVB-Vacuole Fusion. *Current Biology*, *24*(12), 1383-1389. https://doi.org/10.1016/j.cub.2014.05.005

Siol, M., Prosperi, J. M., Bonnin, I., & Ronfort, J. (2008). How multilocus genotypic pattern helps to understand the history of selfing populations : A case study in Medicago truncatula. *Heredity*, *100*(5), 517-525. https://doi.org/10.1038/hdy.2008.5

Siol, M., Wright, S. I., & Barrett, S. C. H. (2010). The population genomics of plant adaptation. *New Phytologist*, *188*(2), 313-332. https://doi.org/10.1111/j.1469-8137.2010.03401.x

Söderström, L., Hagborg, A., Von Konrat, M., Bartholomew-Began, S., Bell, D., Briscoe, L., Brown, E., Cargill, D. C., Da Costa, D. P., Crandall-Stotler, B. J., Cooper, E., Dauphin, G., Engel, J., Feldberg, K., Glenny, D., Gradstein, S. R., He, X., Hentschel, J., Ilkiu-Borges, A. L., … Zhu, R.-L. (2016). World checklist of hornworts and liverworts. *PhytoKeys*, *59*, 1-828. https://doi.org/10.3897/phytokeys.59.6261

Solly, J. E., Cunniffe, N. J., & Harrison, C. J. (2017). Regional Growth Rate Differences Specified by Apical Notch Activities Regulate Liverwort Thallus Shape. *Current Biology*, *27*(1), 16-26. https://doi.org/10.1016/j.cub.2016.10.056

Song, B., Ning, W., Wei, D., Jiang, M., Zhu, K., Wang, X., Edwards, D., Odeny, D. A., & Cheng, S. (2023). Plant genome resequencing and population genomics : Current status and future prospects. *Molecular Plant*, *16*(8), 1252-1268. https://doi.org/10.1016/j.molp.2023.07.009

Song, J.-M., Guan, Z., Hu, J., Guo, C., Yang, Z., Wang, S., Liu, D., Wang, B., Lu, S., Zhou, R., Xie, W.-Z., Cheng, Y., Zhang, Y., Liu, K., Yang, Q.-Y., Chen, L.-L., & Guo, L. (2020). Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of Brassica napus. *Nature Plants*, *6*(1), 34-45. https://doi.org/10.1038/s41477-019-0577-7

Soucy, S. M., Huang, J., & Gogarten, J. P. (2015). Horizontal gene transfer : Building the web of life. *Nature Reviews Genetics*, *16*(8), 472-482. https://doi.org/10.1038/nrg3962

Stahl, E. A., Dwyer, G., Mauricio, R., Kreitman, M., & Bergelson, J. (1999). Dynamics of disease resistance polymorphism at the Rpm1 locus of Arabidopsis. *Nature*, *400*(6745), 667-671. https://doi.org/10.1038/23260

Stephens, M., & Donnelly, P. (2000). Inference in Molecular Population Genetics. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *62*(4), 605-635. https://doi.org/10.1111/1467-9868.00254

Sugano, S. S., Nishihama, R., Shirakawa, M., Takagi, J., Matsuda, Y., Ishida, S., Shimada, T., Hara-Nishimura, I., Osakabe, K., & Kohchi, T. (2018). Efficient CRISPR/Cas9-based genome editing and its application to conditional genetic analysis in Marchantia polymorpha. *PLOS ONE*, *13*(10), e0205117. https://doi.org/10.1371/journal.pone.0205117

Sun, X., Sun, C., Li, Z., Hu, Q., Han, L., & Luo, H. (2016). AsHSP17, a creeping bentgrass small heat shock protein modulates plant photosynthesis and ABA-dependent and independent signalling to attenuate plant response to abiotic stress. *Plant, Cell & Environment*, *39*(6), 1320-1337. https://doi.org/10.1111/pce.12683

Sun, Y., Wang, J., Li, Y., Jiang, B., Wang, X., Xu, W.-H., Wang, Y.-Q., Zhang, P.-T., Zhang, Y.-J., & Kong, X.-D. (2022). Pan-Genome Analysis Reveals the Abundant Gene Presence/Absence Variations Among Different Varieties of Melon and Their Influence on Traits. *Frontiers in Plant Science*, *13*, 835496. https://doi.org/10.3389/fpls.2022.835496

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, *123*(3), 585-595.

Takezawa, D., Komatsu, K., & Sakata, Y. (2011). ABA in bryophytes : How a universal growth regulator in life became a plant hormone? *Journal of Plant Research*, *124*(4), 437-453. https://doi.org/10.1007/s10265-011-0410-5

Takizawa, R., Hatada, M., Moriwaki, Y., Abe, S., Yamashita, Y., Arimitsu, R., Yamato, K. T., Nishihama, R., Kohchi, T., Koeduka, T., Chen, F., & Matsui, K. (2021). Fungal-Type Terpene Synthases in *Marchantia polymorpha* Are Involved in Sesquiterpene Biosynthesis in Oil Body Cells. *Plant and Cell Physiology*, *62*(3), 528-537. https://doi.org/10.1093/pcp/pcaa175

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., & Meyre, D. (2019). Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics*, *20*(8), 467-484. https://doi.org/10.1038/s41576-019-0127-1

Tan, Q. W., Lim, P. K., Chen, Z., Pasha, A., Provart, N., Arend, M., Nikoloski, Z., & Mutwil, M. (2023). Cross-stress gene expression atlas of Marchantia polymorpha reveals the hierarchy and regulatory principles of abiotic stress responses. *Nature Communications*, *14*(1), 986. https://doi.org/10.1038/s41467-023-36517-w

Tao, Y., Luo, H., Xu, J., Cruickshank, A., Zhao, X., Teng, F., Hathorn, A., Wu, X., Liu, Y., Shatte, T., Jordan, D., Jing, H., & Mace, E. (2021). Extensive variation within the pan-genome of cultivated and wild sorghum. *Nature Plants*, *7*(6), 766-773. https://doi.org/10.1038/s41477-021-00925-x

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., Angiuoli, S. V., Crabtree, J., Jones, A. L., Durkin, A. S., DeBoy, R. T., Davidsen, T. M., Mora, M., Scarselli, M., Margarit y Ros, I., Peterson, J. D., Hauser, C. R., Sundaram, J. P., Nelson, W. C., … Fraser, C. M. (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae* : Implications for the microbial

"pan-genome". *Proceedings of the National Academy of Sciences*, *102*(39), 13950-13955. https://doi.org/10.1073/pnas.0506758102

The Angiosperm Phylogeny Group. (2016). An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants : APG IV. *Botanical Journal of the Linnean Society*, *181*(1), 1-20. https://doi.org/10.1111/boj.12385

Tibbs Cortes, L., Zhang, Z., & Yu, J. (2021). Status and prospects of genome-wide association studies in plants. *The Plant Genome*, *14*(1). https://doi.org/10.1002/tpg2.20077

Togninalli, M., Seren, Ü., Meng, D., Fitz, J., Nordborg, M., Weigel, D., Borgwardt, K., Korte, A., & Grimm, D. G. (2018). The AraGWAS Catalog : A curated and standardized Arabidopsis thaliana GWAS catalog. *Nucleic Acids Research*, *46*(D1), D1150-D1156. https://doi.org/10.1093/nar/gkx954

Van de Weyer, A.-L., Monteiro, F., Furzer, O. J., Nishimura, M. T., Cevik, V., Witek, K., Jones, J. D. G., Dangl, J. L., Weigel, D., & Bemm, F. (2019). A Species-Wide Inventory of NLR Genes and Alleles in Arabidopsis thaliana. *Cell*, *178*(5), 1260-1272.e14. https://doi.org/10.1016/j.cell.2019.07.038

Van Holle, S., & Van Damme, E. J. M. (2019). Messages From the Past : New Insights in Plant Lectin Evolution. *Frontiers in Plant Science*, *10*, 36. https://doi.org/10.3389/fpls.2019.00036

Van Zanten, M., Snoek, L. B., Proveniers, M. C. G., & Peeters, A. J. M. (2009). The many functions of ERECTA. *Trends in Plant Science*, *14*(4), 214-218. https://doi.org/10.1016/j.tplants.2009.01.010

Villarreal A., J. C., Crandall-Stotler, B. J., Hart, M. L., Long, D. G., & Forrest, L. L. (2016). Divergence times and the evolution of morphological complexity in an early land plant lineage (Marchantiopsida) with a slow molecular rate. *New Phytologist*, *209*(4), 1734-1746. https://doi.org/10.1111/nph.13716

Vitti, J. J., Grossman, S. R., & Sabeti, P. C. (2013). Detecting Natural Selection in Genomic Data. *Annual Review of Genetics*, *47*(1), 97-120. https://doi.org/10.1146/annurev-genet-111212-133526

Wakeley, J. (2009). *Coalescent theory : An introduction*. Roberts & Co. Publishers.

Wang, L., Wan, M.-C., Liao, R.-Y., Xu, J., Xu, Z.-G., Xue, H.-C., Mai, Y.-X., & Wang, J.-W. (2023). The maturation and aging trajectory of Marchantia polymorpha at single-cell resolution. *Developmental Cell*, *58*(15), 1429-1444.e6. https://doi.org/10.1016/j.devcel.2023.05.014

Weigel, D., & Nordborg, M. (2015). Population Genomics for Understanding Adaptation in Wild Plant Species. *Annual Review of Genetics*, *49*(1), 315-338. https://doi.org/10.1146/annurev-genet-120213-092110

Werner, G. D. A., Cornelissen, J. H. C., Cornwell, W. K., Soudzilovskaia, N. A., Kattge, J., West, S. A., & Kiers, E. T. (2018). Symbiont switching and alternative resource acquisition strategies drive mutualism breakdown. *Proceedings of the National Academy of Sciences*, *115*(20), 5229-5234. https://doi.org/10.1073/pnas.1721629115

Whitton, J. (2013). Plant Biodiversity, Overview. In *Encyclopedia of Biodiversity* (p. 56-64). Elsevier. https://doi.org/10.1016/B978-0-12-384719-5.00110-6

Wolf, Y. I., & Koonin, E. V. (2013). Genome reduction as the dominant mode of evolution. *BioEssays*, *35*(9), 829-837. https://doi.org/10.1002/bies.201300037

Wright, S. (1921). SYSTEMS OF MATING. III. ASSORTATIVE MATING BASED ON SOMATIC RESEMBLANCE. *Genetics*, *6*(2), 144-161. https://doi.org/10.1093/genetics/6.2.144

Wright, S. (1965). The Interpretation of Population Structure by F-Statistics with Special Regard to Systems of Mating. *Evolution*, *19*(3), 395. https://doi.org/10.2307/2406450

Wu, Q., Han, T.-S., Chen, X., Chen, J.-F., Zou, Y.-P., Li, Z.-W., Xu, Y.-C., & Guo, Y.-L. (2017). Long-term balancing selection contributes to adaptation in Arabidopsis and its relatives. *Genome Biology*, *18*(1), 217. https://doi.org/10.1186/s13059-017-1342-8

Xiang, Y., Song, B., Née, G., Kramer, K., Finkemeier, I., & Soppe, W. J. J. (2016). Sequence Polymorphisms at the *REDUCED DORMANCY5* Pseudophosphatase Underlie Natural Variation in Arabidopsis Dormancy. *Plant Physiology*, *171*(4), 2659-2670. https://doi.org/10.1104/pp.16.00525

Xiao, C., Guo, H., Tang, J., Li, J., Yao, X., & Hu, H. (2021). Expression Pattern and Functional Analyses of Arabidopsis Guard Cell-Enriched GDSL Lipases. *Frontiers in Plant Science*, *12*, 748543. https://doi.org/10.3389/fpls.2021.748543

Xu, B., Ohtani, M., Yamaguchi, M., Toyooka, K., Wakazaki, M., Sato, M., Kubo, M., Nakano, Y., Sano, R., Hiwatashi, Y., Murata, T., Kurata, T., Yoneda, A., Kato, K., Hasebe, M., & Demura, T. (2014). Contribution of NAC Transcription Factors to Plant Adaptation to Land. *Science*, *343*(6178), 1505-1508. https://doi.org/10.1126/science.1248417

Xu, B., Taylor, L., Pucker, B., Feng, T., Glover, B. J., & Brockington, S. F. (2021). The land plant-specific MIXTA-MYB lineage is implicated in the early evolution of the plant cuticle and the colonization of land. *New Phytologist*, *229*(4), 2324-2338. https://doi.org/10.1111/nph.16997

Xu, C., Luo, F., & Hochholdinger, F. (2016). LOB Domain Proteins : Beyond Lateral Organ Boundaries. *Trends in Plant Science*, *21*(2), 159-167. https://doi.org/10.1016/j.tplants.2015.10.010

Yamato, K. T., Ishizaki, K., Fujisawa, M., Okada, S., Nakayama, S., Fujishita, M., Bando, H., Yodoya, K., Hayashi, K., Bando, T., Hasumi, A., Nishio, T., Sakata, R., Yamamoto, M., Yamaki, A., Kajikawa, M., Yamano, T., Nishide, T., Choi, S.-H., … Ohyama, K. (2007). Gene organization of the liverwort Y chromosome reveals distinct sex chromosome evolution in a haploid system. *Proceedings of the National Academy of Sciences*, *104*(15), 6472-6477. https://doi.org/10.1073/pnas.0609054104

Yang, L., Wang, Z., & Hua, J. (2021). A Meta-Analysis Reveals Opposite Effects of Biotic and Abiotic Stresses on Transcript Levels of Arabidopsis Intracellular Immune Receptor Genes. *Frontiers in Plant Science*, *12*, 625729. https://doi.org/10.3389/fpls.2021.625729

Yang, T.-H., Lenglet-Hilfiker, A., Stolz, S., Glauser, G., & Farmer, E. E. (2020). Jasmonate Precursor Biosynthetic Enzymes LOX3 and LOX4 Control Wound-Response Growth Restriction. *Plant Physiology*, *184*(2), 1172-1180. https://doi.org/10.1104/pp.20.00471

Yang, X., Lee, W.-P., Ye, K., & Lee, C. (2019). One reference genome is not enough. *Genome Biology*, *20*(1), 104. https://doi.org/10.1186/s13059-019-1717-0

Yoder, J. B., Briskine, R., Mudge, J., Farmer, A., Paape, T., Steele, K., Weiblen, G. D., Bharti, A. K., Zhou, P., May, G. D., Young, N. D., & Tiffin, P. (2013). Phylogenetic Signal Variation in the Genomes of Medicago (Fabaceae). *Systematic Biology*, *62*(3), 424-438. https://doi.org/10.1093/sysbio/syt009

Yokota, T., Ohnishi, T., Shibata, K., Asahina, M., Nomura, T., Fujita, T., Ishizaki, K., & Kohchi, T. (2017). Occurrence of brassinosteroids in non-flowering land plants, liverwort, moss, lycophyte and fern. *Phytochemistry*, *136*, 46-55. https://doi.org/10.1016/j.phytochem.2016.12.020

Yu, J., Cai, Y., Zhu, Y., Zeng, Y., Dong, S., Zhang, K., Wang, S., Li, L., Goffinet, B., Liu, H., & Liu, Y. (2022). Chromosome-Level Genome Assemblies of Two Hypnales (Mosses) Reveal High Intergeneric Synteny. *Genome Biology and Evolution*, *14*(2), evac020. https://doi.org/10.1093/gbe/evac020

Yu, J., Golicz, A. A., Lu, K., Dossa, K., Zhang, Y., Chen, J., Wang, L., You, J., Fan, D., Edwards, D., & Zhang, X. (2019). Insight into the evolution and functional characteristics of the pan-genome assembly from sesame landraces and modern cultivars. *Plant Biotechnology Journal*, *17*(5), 881-892. https://doi.org/10.1111/pbi.13022

Yu, J., Li, L., Wang, S., Dong, S., Chen, Z., Patel, N., Goffinet, B., Chen, H., Liu, H., & Liu, Y. (2020). Draft genome of the aquatic moss Fontinalis antipyretica (Fontinalaceae, Bryophyta). *Gigabyte*, *2020*, 1-9. https://doi.org/10.46471/gigabyte.8

Yuan, Y., Bayer, P. E., Batley, J., & Edwards, D. (2021). Current status of structural variation studies in plants. *Plant Biotechnology Journal*, *19*(11), 2153-2163. https://doi.org/10.1111/pbi.13646

Zeng, K., Fu, Y.-X., Shi, S., & Wu, C.-I. (2006). Statistical Tests for Detecting Positive Selection by Utilizing High-Frequency Variants. *Genetics*, *174*(3), 1431-1439. https://doi.org/10.1534/genetics.106.061432

Zhang, C., Dong, S.-S., Xu, J.-Y., He, W.-M., & Yang, T.-L. (2019). PopLDdecay : A fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics*, *35*(10), 1786-1788. https://doi.org/10.1093/bioinformatics/bty875

Zhang, J., Fu, X.-X., Li, R.-Q., Zhao, X., Liu, Y., Li, M.-H., Zwaenepoel, A., Ma, H., Goffinet, B., Guan, Y.-L., Xue, J.-Y., Liao, Y.-Y., Wang, Q.-F., Wang, Q.-H., Wang, J.-Y., Zhang, G.-Q., Wang, Z.-W., Jia, Y., Wang, M.-Z., … Chen, Z.-D. (2020). The hornwort genome and early land plant evolution. *Nature Plants*, *6*(2), 107-118. https://doi.org/10.1038/s41477-019-0588-4

Zhang, J., & Sun, X. (2021). Recent advances in polyphenol oxidase-mediated plant stress responses. *Phytochemistry*, *181*, 112588. https://doi.org/10.1016/j.phytochem.2020.112588

Zhang, Z., Ersoz, E., Lai, C.-Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., Bradbury, P. J., Yu, J., Arnett, D. K., Ordovas, J. M., & Buckler, E. S. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics*, *42*(4), 355-360. https://doi.org/10.1038/ng.546

Zhao, J., Bayer, P. E., Ruperao, P., Saxena, R. K., Khan, A. W., Golicz, A. A., Nguyen, H. T., Batley, J., Edwards, D., & Varshney, R. K. (2020). Trait associations in the pangenome of pigeon pea (*Cajanus cajan*). *Plant Biotechnology Journal*, *18*(9), 1946-1954. https://doi.org/10.1111/pbi.13354

Zhao, Q., Feng, Q., Lu, H., Li, Y., Wang, A., Tian, Q., Zhan, Q., Lu, Y., Zhang, L., Huang, T., Wang, Y., Fan, D., Zhao, Y., Wang, Z., Zhou, C., Chen, J., Zhu, C., Li, W., Weng, Q., … Huang, X. (2018). Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nature Genetics*, *50*(2), 278-284. https://doi.org/10.1038/s41588-018-0041-z

Zhao, Q., Leung, S., Corbett, A. H., & Meier, I. (2006). Identification and Characterization of the Arabidopsis Orthologs of Nuclear Transport Factor 2, the Nuclear Import Factor of Ran. *Plant Physiology*, *140*(3), 869-878. https://doi.org/10.1104/pp.105.075499

Zhao, Y.-P., Fan, G., Yin, P.-P., Sun, S., Li, N., Hong, X., Hu, G., Zhang, H., Zhang, F.-M., Han, J.-D., Hao, Y.-J., Xu, Q., Yang, X., Xia, W., Chen, W., Lin, H.-Y., Zhang, R., Chen, J., Zheng, X.-M., … Ge, S. (2019). Resequencing 545 ginkgo genomes across the world reveals the evolutionary history of the living fossil. *Nature Communications*, *10*(1), 4201. https://doi.org/10.1038/s41467-019-12133-5

Zheng, B., Bai, Q., Wu, L., Liu, H., Liu, Y., Xu, W., Li, G., Ren, H., She, X., & Wu, G. (2019). EMS1 and BRI1 control separate biological processes via extracellular domain diversity and intracellular domain conservation. *Nature Communications*, *10*(1), 4165. https://doi.org/10.1038/s41467-019-12112-w

Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., & Weir, B. S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*, *28*(24), 3326-3328. https://doi.org/10.1093/bioinformatics/bts606

Zhou, P., Silverstein, K. A. T., Ramaraj, T., Guhlin, J., Denny, R., Liu, J., Farmer, A. D., Steele, K. P., Stupar, R. M., Miller, J. R., Tiffin, P., Mudge, J., & Young, N. D. (2017). Exploring structural variation and gene family architecture with De Novo assemblies of 15 Medicago genomes. *BMC Genomics*, *18*(1), 261. https://doi.org/10.1186/s12864-017-3654-1

Zhou, X., Liao, H., Chern, M., Yin, J., Chen, Y., Wang, J., Zhu, X., Chen, Z., Yuan, C., Zhao, W., Wang, J., Li, W., He, M., Ma, B., Wang, J., Qin, P., Chen, W., Wang, Y., Liu, J., … Chen, X. (2018). Loss of function of a rice TPR-domain RNA-binding protein confers broad-spectrum disease resistance. *Proceedings of the National Academy of Sciences*, *115*(12), 3174-3179. https://doi.org/10.1073/pnas.1705927115

Zhou, X., & Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*, *44*(7), 821-824. https://doi.org/10.1038/ng.2310

Zhou, Y., Zhang, Z., Bao, Z., Li, H., Lyu, Y., Zan, Y., Wu, Y., Cheng, L., Fang, Y., Wu, K., Zhang, J., Lyu, H., Lin, T., Gao, Q., Saha, S., Mueller, L., Fei, Z., Städler, T., Xu, S., … Huang, S. (2022). Graph pangenome captures missing heritability and empowers tomato breeding. *Nature*, *606*(7914), 527-534. https://doi.org/10.1038/s41586-022-04808-9

Zou, J., Liu, C., Liu, A., Zou, D., & Chen, X. (2012). Overexpression of OsHsp17.0 and OsHsp23.7 enhances drought and salt tolerance in rice. *Journal of Plant Physiology*, *169*(6), 628-635. https://doi.org/10.1016/j.jplph.2011.12.014

Zschiesche, W., Barth, O., Daniel, K., Böhme, S., Rausche, J., & Humbeck, K. (2015). The zinc-binding nuclear protein HIPP 3 acts as an upstream regulator of the salicylate-dependent plant immunity pathway and of flowering time in *Arabidopsis thaliana*. *New Phytologist*, *207*(4), 1084-1096. https://doi.org/10.1111/nph.13419