Improvise. Adapt. Overcome.

René Boesten

Propositions

1. Findings resulting from genome-wide association analyses are speculative until experimentally validated.
   (this thesis)

2. Environmental stress accelerates adaptive evolution.
   (this thesis)

3. The claim that improving photosynthesis will increase crop yields is based more on wishful thinking than on scientific evidence.

4. Conservation biologists should abandon their nativist world views.

5. Open access AI will benefit education by shifting the emphasis towards critical thinking, rather than training 'monkeys' and 'parrots'.

6. Any species of animal, besides humans, are suitable as a pet.

7. Fully paved gardens should be prohibited.

Propositions belonging to the thesis, entitled

Improvise. Adapt. Overcome.

René Boesten
Wageningen, 21 February 2024

Improvise. Adapt. Overcome.

René Boesten

Thesis committee

Promotor
Prof. Dr Mark G.M. Aarts
Personal chair at the Laboratory of Genetics
Wageningen University & Research

Co-promotor
Prof. Dr Bas J. Zwaan
Professor of Genetics
Wageningen University & Research

Other members
Prof. Dr Viola A. Willemsen, Wageningen University & Research
Prof. Dr Juliette de Meaux, University of Cologne, Germany
Dr Fabrice Roux, LIPME, INRAE, CNRS, University of Toulouse, France
Prof. Dr Stephan Clemens, University of Bayreuth, Germany

This research was conducted under the auspices of the Graduate School Experimental Plant Sciences.

Improvise. Adapt. Overcome.

René Boesten

Thesis
Submitted in fulfillment of the requirements for the degree of doctor
at Wageningen University,
by the authority of the Rector Magnificus,
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Wednesday 21 February 2024
at 4 p.m. in the Omnia Auditorium.

René Boesten

Improvise. Adapt. Overcome., 208 pages

PhD thesis, Wageningen University, the Netherlands (2024)
With references, with summary in English

DOI: https://doi.org/10.18174/646283

Table of contents

Chapter 1

General introduction

Plant adaptation to their environment

The environment of a plant encompasses numerous biotic and abiotic factors that fluctuate over time. Once these negatively impact plant performance or fitness, they can be considered as sources of (a)biotic stress. Herbivory, pathogens, and competing plants are common causes of biotic stress (Suzuki *et al*., 2014), and water (drought or flooding), salinity, extreme temperatures, light levels, nutrient deficiencies and excesses are examples of frequently occurring sources of abiotic stress (Pereira, 2016). It is important to note that the use of the term 'stress' is context depended as conditions that negatively impact one species, population or individual, may be optimal to another. As sessile organisms, plants cannot evade most sources of stress. The reproductive fitness of an individual plant is thus largely dependent on its capability to adapt its responses to best match their environment (Figure 1).

Broadly speaking, plants can use adaptive phenotypic plasticity to deal with immediate stresses that act upon a single generation. Over multiple generations, plants can adjust their genotype via adaptive evolution to better match their environment or can move, for instance via seed dispersal, to better suited environments. The adaptative and evolutionary principles of plant adaptation that are at play in a natural setting are also of major importance to agriculture. On the one hand, crops may suffer major yield losses resulting from different kinds of (a)biotic stresses. On the other hand, the environment sets the borders in which a particular species can grow and thrive. By developing better-adapted and more resilient crops,

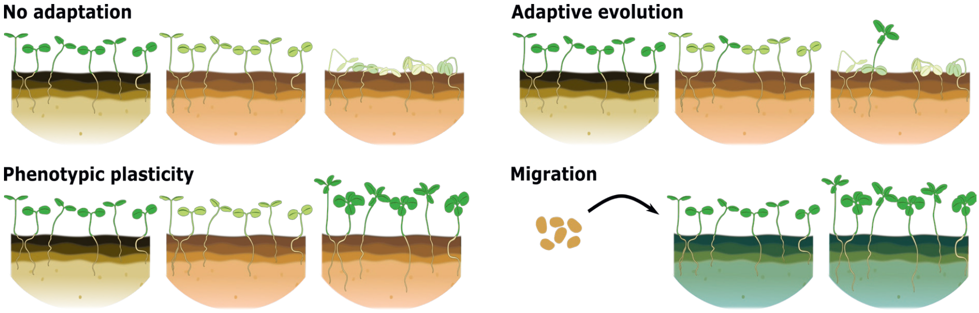


Figure 1: **Plants may adapt through different mechanisms to stress.** When exposed to stress (indicated with a reddish soil) and no appropriate mechanisms to deal with it are present, plants may perish (top left panel). Alternatively, plants may adjust their physiology and morphology is such a way to better cope with the particular stress, which is known as adaptive phenotypic plasticity (bottom left panel). There may be natural genetic variation for the response to stress, where only individuals with a particular genotype, that is better suited to deal with the environment, may survive (top right panel). Such genotypes can be selected for from standing genetic variation, or may arise through spontaneous mutation. Finally, if a genotype cannot survive in a particular environment, it may migrate to environments to which it is better suited, for instance via seed dispersal (bottom right panel).

these losses can be mitigated and the environmental compatibility of a crop can be expanded. Consequently, the total suitable area for cultivation, as well as the yield per unit area, can be increased. Therefore, it is crucial to comprehend the mechanisms of adaptation available to plants.

Shapeshifting as a means to succeed

Throughout a plant's lifespan, adaptation to the environment occurs primarily via adaptive phenotypic plasticity. This refers to the organism's capacity to alter its physiology or morphology in response to changes in environmental conditions in ways that enhance the organisms performance and fitness (Schlichting, 1986). Phenotypic plasticity encompasses various processes, starting from the initial perception of stress, followed by signalling, signal integration, and ultimately culminating in a response achieved through adjustments in transcriptional and translational processes (H. Zhang *et al*., 2022). Different stressors often elicit specific and distinct responses tailored to match those specific stressors. However, multiple stresses often co-occur in nature, and they may individually induce different or even opposing responses. For instance, photosynthesis requires open stomata, which leads to high transpiration, to function properly. Therefore, growth is not possible when stomata are closed. Heat stress prompts plants to open their stomata, thereby increasing transpiration to cool the leaves. Conversely, drought stress leads plants to close their stomata to prevent transpiration and water loss (Rivero *et al*., 2022). When heat and drought occur together, which is common, their combined effect cannot simply be the sum of the two individual responses. Instead, it requires prioritization or induction of unique responses (Mittler, 2006; Rivero *et al*., 2022; Suzuki *et al*., 2014; Thoen *et al*., 2017).

In addition to dealing with immediate stresses, plants appear to possess the ability to utilize environmental cues to anticipate future stresses, which further contributes to the complexity of their plastic responses. The likely existence of anticipatory adaptative mechanisms in plants is not surprising, considering that certain stresses often follow a correlated sequential order in nature (Mertens, Boege, *et al*., 2021). For instance, subtle decreases in day-length and temperature near the end of the growth season will tightly correlate with further and more stressful decreases of temperature that occur during winter. Moreover, the arrival of insect herbivores during a season typically adheres to common sequences of arrival of insects that are either leaf-chewing (e.g., caterpillars) or phloem-feeding (e.g., aphids). The defence response against leaf-chewing insects is predominantly regulated through jasmonic acid (JA) signalling, while the response to phloem-feeding insects is primarily regulated through salicylic acid (SA) signalling (J. Wu & Baldwin, 2010). The defence response against either leaf-chewing or phloem-feeding insects generally enhances tolerance to subsequent attacks from the same feeding type. However, it may also increase susceptibility to subsequent attacks from the opposing feeding type due to the antagonistic SA-JA crosstalk (Soler *et al*., 2012; Thaler *et al*., 2012). Nevertheless, a field study conducted by (Mertens, Fernández de Bobadilla, *et al*., 2021) demonstrates that despite the antagonistic responses to these two types of herbivores, plants retain their resistance to consecutive herbivore attacks (of both feeding types) when attackers arrive in a common sequential order. On the other hand, increased susceptibility is observed when rare sequential orders of attack by the same herbivores occurs. Hence, signal transduction and integration of both acute and anticipated stresses thus play a crucial role in the decision-making process of plastic responses.

The initial stages of acclimation involve sensing and signalling environmental changes. Early stress sensing primarily occurs through physical and chemical alterations that arise as a direct consequence of environmental shifts (Baxter *et al*., 2014). These early sensing mechanisms subsequently initiate cellular responses through a range of early signalling events, such as the activation of mitogen-activated protein kinases (MAPKs), protein phosphorylation and various second messengers including Ca^{2+}, phospholipids, nitric oxide and reactive oxygen species (ROS) such as superoxide anion, hydrogen peroxide, hydroxyl radical and singlet oxygen (Baxter *et al*., 2014; H. Zhang *et al*., 2022). Notably, ROS play an interesting dual role. They serve as crucial signalling molecules required for an effective stress response; however, when their levels exceed the plant's detoxification capacity, they can cause irreversible oxidative damage to DNA, RNA, proteins, lipids and numerous small molecules (Miller *et al*., 2009). Since ROS also play a vital role in various developmental processes such as plant organ morphogenesis, it is essential to maintain a balance in their production and quenching (H. Huang *et al*., 2019). This balance is primarily achieved by compartmentalizing ROS production, regulating its generation in a spatial and temporal manner and controlling ROS scavenging agents (Castro *et al*., 2021).

The ability of plants to adjust their physiology to better suit the environment can be influenced by past experiences during their lifetime. This phenomenon, known as stress priming, occurs when exposure to a specific type of (a)biotic stress earlier in life affects the response to subsequent stress later on, often rendering the plant more resistant to the latter stress (Bruce *et al*., 2007; H. Liu *et al*., 2022). For instance, wheat plants whose seeds were pretreated by soaking them in saline solution displayed increased tolerance to salinity throughout their growing season (Iqbal & Ashraf, 2007). The memory associated with stress priming is thought to be regulated epigenetically through DNA methylation, chromatin remodeling, histone modifications and small RNAs (Turgut-Kara *et al*., 2020). Generally, these epigenetic marks function in gene expression regulation and influence genome stability by silencing transposable elements (TEs) (Law & Jacobsen, 2010). The duration of these epigenetic modifications can vary, ranging from hours to weeks, with some even being transmitted to subsequent generations, potentially serving as transgenerational stress memory (Boyko *et al*., 2006; Mladenov *et al*., 2021). However, it is important to note that only some epigenetic modifications are adaptive and the majority of epigenetic modifications are erased during meiosis. Those that persist typically only endure for a single subsequent generation (Mladenov *et al*., 2021). As a result, no epi-alleles have been reported in plants that confer persistent stress tolerance after the initial application of stress, although the evolution of such epi-alleles would be expected if such memory existed (H. Zhang *et al*., 2022). Epigenetic stress memory may be particularly important for asexually reproducing plants, as they bypass the epigenetic bottleneck of meiosis.

Shaping success

The second primary mechanism for adaptation is the acquisition of beneficial genetic variants through the process of adaptive evolution. Three requirements must be met for adaptive evolution to occur: (i) there must be variation in traits, (ii) the variation in traits must be associated with variable fitness (i.e., selection) and (iii) there must be heritability (parent-offspring resemblance) of the trait variation. The factors together can result in an increase in the frequency of beneficial alleles (associated with higher fitness) at the expense of deleterious alleles in a population over successive generations. Since the environment provides the context for determining what is advantageous or deleterious, in populations that inhabit different environments, different alleles may accumulate that confer benefits specific to those environments.

Types of genetic variants

Heritable trait variation is thus crucial for facilitating adaptive evolution. Variation can arise through segregation and recombination, where existing genetic variants in parental genomes are reshuffled in offspring, or through the occurrence of spontaneous mutations. These mutations can be classified based on their size, including single nucleotide polymorphisms (SNPs; 1 bp), small insertions and deletions (indels; 2– 50 bp), structural variations (SVs; >50 bp) such as copy number variants (CNVs), inversions and translocations, or even encompass entire chromosomes in the case of aneuploidy or altered ploidy levels (Figure 2a). Although the classification of variant types based on size is somewhat arbitrary, their relative contributions to (adaptive) evolution may vary (Figure 2b). The underlying causes of spontaneous mutations may differ for these different types of variants. Most variant types arise from common mechanisms and causes, such as DNA replication errors, faulty DNA repair, activation of mobile genetic elements, recombination errors during mitosis or meiosis and exposure to mutagenic factors such as radiation and certain chemicals. However, the predominant causes and mechanisms differ between variant types and often depend on the genomic context. For instance, at CpG sites (CG dinucleotides where both cytosines are methylated), the methylated cytosines are prone to being converted to thymine, leading to a SNP if not corrected by mismatch repair (G.-M. Li, 2008). Indels may result from strand slippage of the polymerase during DNA replication, which occurs more frequently in repetitive sequences such as microsatellites (Campbell & Eichler, 2013; G.-M. Li, 2008). CNVs are more likely to occur in genomic regions with low-copy repeats that facilitate nonallelic homologous recombination (Zhang *et al*., 2009; Żmieńko *et al*., 2014). Therefore, both the genomic context in which spontaneous mutations are more likely to arise and the rate which they occur differ for the different variant types.

The dynamic nature of mutation rates

Currently, there is still much that remains unknown about spontaneous mutation rates in plants in general, let alone the rates for each specific variant type. Several biological and technical challenges make it difficult to accurately measure mutation rates (Quiroz *et al*., 2023). One of the reasons is that the absolute number of spontaneous mutations per generation is very low. Early experiments aimed to measure mutation rates by observing *de novo* phenotype generation for easily detectable traits. For example, in maize, the *R* gene was used as a phenotypic marker since mutations in this gene would lead to colourless plants and/or colourless kernel aleurone (Stadler, 1946). However, using phenotypes as a marker to estimate mutation rates has drawbacks, as it requires screening a large number of progeny due to the low occurrence of spontaneous mutations. Additionally, the resulting observed mutation rates heavily depend on the genetic complexity of the trait under study. For instance, if a change in phenotype is generated by a loss-of-function mutation in a specific gene, there can be many different mutations (e.g., any frameshift mutations) that result in the same phenotype, compared to traits obtained through more specific gain-of-function alleles. Furthermore, the obtained estimates only provide a gene- or trait-specific mutation rate, which may not accurately represent the genome-wide mutation rates, and they do not differentiate between the different types of variants that cause the phenotypic changes.

Nowadays, the increasing availability of next-generation sequencing methods has made it possible to more directly assess mutation rates. One approach is to sequence both parents and offspring to identify mutations present in the offspring but not in either parent. Another common method is to estimate mutation rates through sequencing of individuals from mutation accumulation experiments. In mutation accumulation experiments, offspring of a single, fully homozygous parental line are grown under optimal conditions and are propagated through single-seed descent for multiple generations (Halligan & Keightley, 2009; Ossowski *et al*., 2010; Quiroz *et al*., 2023). The optimal experimental conditions aim to minimize selection pressure, while propagation through single-seed descent creates a genetic bottleneck, allowing nearly all spontaneous mutations to accumulate (Halligan & Keightley, 2009). The use of genetic bottlenecks and minimal selection is important because otherwise deleterious mutations could be rapidly selected against and lost from the experimental lines, leading to an underestimation of the total number of mutations that occurred. Based on such experiments, current genome-wide mutation rate estimates for plants range roughly between 0.1 and 1 mutation per generation (Ossowski *et al*., 2010; Schoen & Schultz, 2019). These estimates suggest relatively low mutation rates, particularly considering that they are obtained with minimal selection and would likely be even lower with additional selection pressures. These findings together indicate that the process of adaptation through the generation of spontaneous mutations is slow.

Mutation rates themselves are not fixed and can be subject to selection and change over time. As the rate at which new variation arises can directly impact the rate of adaptive evolution, it is important to understand the

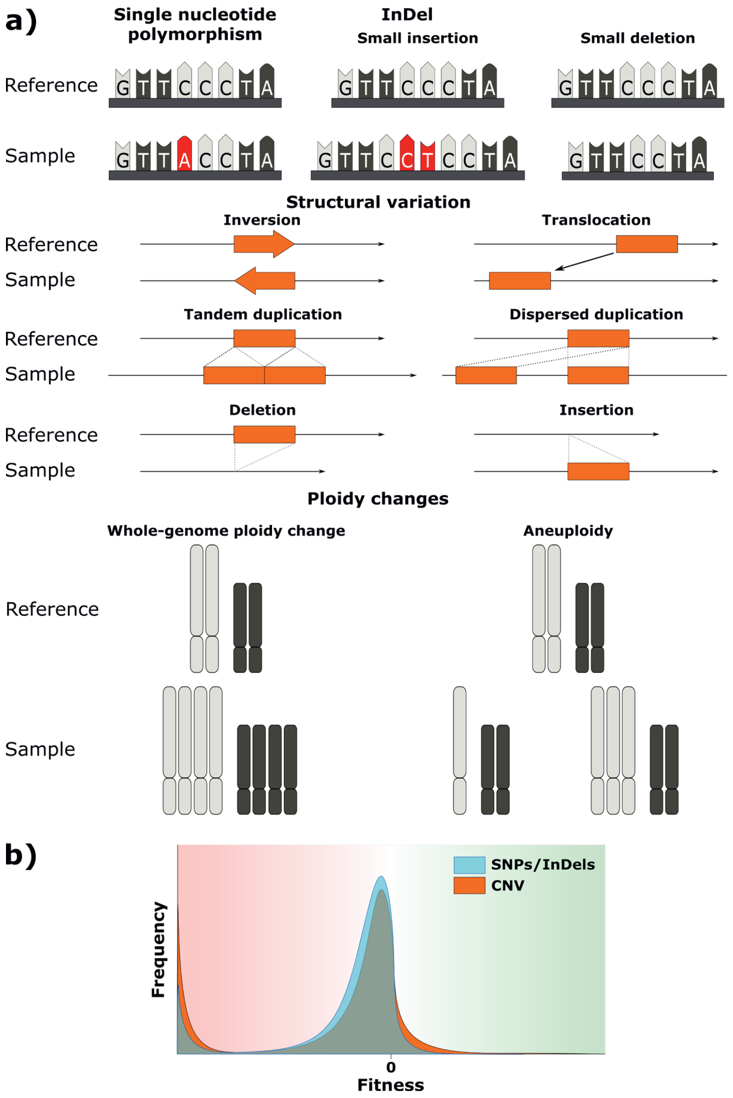


Figure 2: **Mutations range in various shapes and sizes.** a) Single nucleotide polymorphisms (SNPs) span, by definition, only one base-pair (bp). Small insertions and deletions (indels) span 2 – 50 bp, while larger mutations are considered as structural variation, and comprise inversions, translocations and larger gains or losses of sequences. Such larger insertions and deletions, along with gene duplications (either tandem or dispersed) are known as copy number variation. Finally, mutations may also span the entire genome, thus changing the ploidy level, or an entire chromosome(s) in case of aneuploidy. b) Hypothetical impact of mutation size on the relative fitness effect, where larger mutations such as CNVs may more often have larger impacts on fitness compared to smaller mutations such as SNPs and InDels. Note, although these distributions are speculative, the general shape of the fitness effects of mutations is based on the models by Orr, (1998) and a model proposed by Bank *et al*., (2014).

processes that shape mutation rates. For instance, mutations can arise from faulty DNA repair and replication. If the genes responsible for these process themselves mutate, this can lead to a significant increase in the mutation rate. In cases where a population is poorly adapted to its environment, hypermutator phenotypes, with significantly elevated mutation rates, can be selected for. A well-known example of this phenomenon is the evolution of hypermutator phenotypes in several experimental populations of *E. coli*, where mutations in DNA repair mechanism genes resulted in mutation rates up to three orders of magnitude higher than the ancestor (Sniegowski *et al*., 1997; Barrick & Lenski, 2013). Although most mutations are more likely to be deleterious than beneficial, an increased mutation rate also raises the probability of acquiring beneficial mutation. In plants such hyper-mutator alleles are not likely playing much importance, as sexual reproduction will quickly uncouple it from any beneficial mutations. Yet, as long as there is sufficient room for adaptation in the environment, the advantages gained from an increased mutation rate in general may outweigh the burden of the higher load of deleterious mutations. This balance shifts when a population becomes well-adapted, as most mutations will then be deleterious. Therefore, mutation rates are likely in a constant balance between the ability to adapt (higher mutation rate) and the ability to maintain adapted (lower mutation rate) (Ram & Hadany, 2014).

Stress impacts the rate of adaptive evolution

Stressful environments can have a direct impact on mutation rates, providing another avenue for understanding the dynamics of plant adaptation. Stressors such as ROS and alkylating agents can cause mutations either by directly damaging DNA or by indirectly inhibiting the function of enzymes involved in DNA fidelity (Friedberg *et al*., 2005). The specific role of stress-induced mutations in plant adaptation is not yet well understood, but there are indications that this may be an important mechanism aiding in rapid adaptation. A study by (DeBolt, 2010) shed light on this topic using *Arabidopsis thaliana* exposed to different temperatures (16 or 28 ^{o}C) and salicylic acid sprays (mimicking biotic stress) for five consecutive generations while selecting for fecundity. The rapid induction of CNVs by stress suggests that the processes leading to these mutations are susceptible to environmental stress. Two main mechanisms contribute to *de novo* CNVs: non-allelic homologous recombination and activity of transposable elements (TEs). Various (a)biotic stresses have been shown to increase homologous recombination frequency in plants, which likely also enhance the occurrence of non-allelic homologous recombination and directly contribute to de novo CNVs (Boyko *et al*., 2006; Lucht *et al*., 2002; Yao & Kovalchuk, 2011). Additionally, TE activity is increased in response to environmental stresses, partly because certain types of TEs contain stress-responsive cis-elements and because the epigenetic machinery that normally silence TE activity may be compromised (Grandbastien, 2015; Ito *et al*., 2011; Lanciano & Mirouze, 2018). Taken together, these findings suggest that stressful environments may promote the generation of genetic variation in the form of CNVs at an accelerated pace. However, further research is needed to fully understand the implications of stress-induced mutations for plant adaptation to their environments.

CNV, particularly gene duplications, are interesting in the context of rapid plant adaptation as they may result in quick gain-of-function mutations. Higher gene copy numbers can lead to increased gene expression levels and subsequent protein abundance. A notable example illustrating this is the rapid adaptation of plants to the herbicide glyphosate in a natural setting. Glyphosate targets and inhibits the 5-enolypyruvylshikimate-3-phosphate synthase (EPSPS) protein, a crucial component of the Shikimate pathway required for the biosynthesis of several essential amino acids (Steinrücken & Amrhein, 1980). At first, glyphosate tolerance was documented to have rapidly evolved through *EPSPS* gene duplications in *Amaranthus palmeri* (Gaines *et al*., 2010, 2011; Patterson *et al*., 2018). The increased *EPSPS* copy numbers result in elevated gene expression levels and higher protein abundance. As a result, even after glyphosate application, there is enough functional EPSPS protein to sustain the production of vital amino acids. This strategy of rapid adaptation to glyphosate through the acquisition of higher *EPSPS* copy numbers evolved in at least seven other plant species (Patterson *et al*., 2018). This demonstrates that such mutations can strongly impact plant fitness and rapidly spread in highly selective environments. Building upon the study by DeBolt, 2010, these observations further suggest that a stressful environment directly influences the amount of spontaneous genetic variation that arises and simultaneously provides selective pressure on the newly generated variation, ultimately impacting the rate of adaptive evolution. Furthermore, this raises the question of whether the generation of de novo CNVs disproportionally contributes to rapid adaptation to stressful environments.

The challenges of CNV detection

The understanding of the role of CNV in plant adaptation is limited, largely due to technical challenges associated with the detection of this type of genetic variation. Until recently, array-based methods such as array comparative genomic hybridization and SNP arrays were commonly used to identify intraspecific CNVs in plants (Żmieńko *et al*., 2014). While these methods are cost-effective for genotyping large populations, they have several limitations. Arrays provide no positional information on duplications, are not highly sensitive to single copy gains and are better suited for detecting deletions (Alkan *et al*., 2011). Additionally, array-based methods can only detect sequences present in the reference sample used to generate the probes for the array (Kidd *et al*., 2010). The advent of next-generation sequencing (NGS) technologies, particularly short-read Illumina sequencing, has revolutionized the study of CNVs in plants and opened up new possibilities for investigating their contribution to plant adaptation in greater detail. At the start of the research presented in this thesis, there were no bioinformatic tools specifically optimized to study CNV in plant genomes. To address this, close collaboration with the Bioinformatics department, particularly Dr. Raúl Wijfjes, was essential. During the project, he developed a bioinformatics tool called 'Hecaton' with the specific purpose of reliably detecting CNVs in plant genomes from short-read sequencing datasets (Wijfjes *et al*., 2019). Nevertheless, detecting CNVs using short-read sequencing methods remains a technical challenge. More recent developments in long-read sequencing technologies have overcome this challenge to a large extent. The use of long-read sequencing is however not (yet) very feasible to apply on a relatively large sample size. Therefore, the collaboration with the Bioinformatics department allowed for the generation of genetic resources consisting of large sample sizes, their evaluation for CNV prevalence, and attempts to assess the relevance of CNV to plant adaptation.

Outline of this thesis

The central theme of this thesis is to investigate the genetic aspects of plant adaptive responses to their environment, with a special focus on the role of CNV. To this purpose, I study various plant adaptive traits and evaluate what types of genetic variation underlie these traits. Additionally, I will test whether severe abiotic stresses increase the mutation rate of *A. thaliana* and thereby accelerate the generation of adaptive phenotypes. The thesis is organized into chapters that address different aspects of this research topic.

In Chapter 2, I review the current understanding of CNV in plants to better understand its significance to plant evolution and adaptation to the environment. The review highlights that CNV is a prevalent, yet understudied type of genetic variation within and between plant species. I discuss that CNV can have strong immediate impacts on traits, for instance by strongly affecting gene expression levels, while over longer evolutionary timescales allows for the evolution of completely novel genes and traits. Additionally, I discuss what kind of genetic mechanisms give rise to CNV and what the status and challenges are of CNV detection. Finally, I provide various examples of how CNV has played a key role in the evolution of plants that are for instance adapted to more extreme environments and discuss that CNV also affects many agronomically relevant traits.

In Chapter 3, I test how stressful environments impact the rate of adaptive evolution in plants. An experimental evolution approach is employed, which involves studying evolutionary changes in experimental populations under controlled conditions imposed by the experimenter (Kawecki *et al*., 2012). Specifically, I test whether *A. thaliana* can adapt rapidly to a stressful and highly selective environment created by severe salinity and zinc stress. By using an isogenic population of *A. thaliana*, spontaneous genetic and epigenetic mutations that occurred during the experiment are tracked and their impact on plant performance is evaluated. I find that severe stress can induce the mutation rate and that one of the experimental populations exposed to severe zinc stress has drastically improved its performance via a spontaneous genetic mutation within only five generations.

Chapter 4 adopts a different approach and investigates patterns of local adaptation in *A. thaliana* in the Netherlands. A novel natural variation panel is developed, comprising of 192 fully sequenced natural accessions. The natural variation panel demonstrates a high genetic diversity despite that all plants originate from a relatively small geographical area. By using a dense sampling approach, I could examine patterns of local adaptation despite the relatively uniform climate in the Netherlands. For instance, by using genome-environment association analysis I test whether plants have adapted to a subtle cline in climatic variables related to temperature and precipitation. I find a QTL associated with this cline and demonstrate that allelic variation at this QTL affects the response to drought. Additionally, I show that semidwarf accessions, which occur relatively frequently in more windy areas of the countries, are more tolerant to high wind velocities and report a new gene involved in the response to iron deficiency. Together, these examples show that this natural variation panel is suitable to study the genetic basis of plant adaptive traits.

In Chapter 5, the focus shifts to *Hirschfeldia incana*, a member of the Brassicaceae family with unusually high levels of photosynthesis for a C_3-photosynthesizing plant species. Similar to all members of the Brassiceae tribe, this species underwent an ancient whole genome triplication event, followed by extensive genome rearrangements and gene loss during the process of returning to diploidy. Different species within the Brassiceae tribe have experienced variations in gene retention and loss, leading to divergence among species and potentially contributed to unique traits and characteristics. The hypothesis put forward is that *H. incana* has retained higher copy numbers of genes involved in photosynthesis, which may have contributed to its elevated photosynthesis capacity. To test this hypothesis, a reference genome of *H. incana* is generated, and a comparative genomics approach is utilized to evaluate the impact of gene duplications associated to photosynthesis. In addition, gene expression levels are measured for a subset of photosynthesis-related genes that are duplicated in *H. incana*, to further elucidate the role of CNVs in the unique photosynthetic characteristics of this species.

Finally, in Chapter 6, the main topics and results of the thesis are discussed in a broader context, providing a comprehensive overview of the research findings and their implications to gain further insights into the adaptive responses of plants to their environment.

Chapter 2

The role of copy number variation in the evolution of plants

René Boesten[1*], Raúl Wijfjes[2,3*], Sandra Smit[2], Dick de Ridder[2] and Mark G. M. Aarts[1]

[1] Laboratory of Genetics, Wageningen University & Research
[2] Bioinformatics Group, Wageningen University & Research
[3] Current affiliation: Faculty of Biology, Ludwig Maximilian University of Munich

[*]These authors contributed equally to this work.

Abstract

Copy number variation (CNV), a form of intraspecific genetic variation involving gain or loss of genomic DNA fragments, typically in the 50 basepairs (bp) to 100 kilobp (kbp) range, is a major component underlying phenotypic variation in many organisms. CNVs are abundant among plants and their generation appears to be highly dynamic, altering plant genome sequences on a short evolutionary time scale. Initially, CNVs involving gene duplications are genetically mostly redundant, largely affecting expression of the duplicated gene. Eventually, duplicated genes may evolve altered or new functions, and lose their redundancy. CNVs are commonly linked to adaptation and evolution of plants in response to suboptimal or adverse environmental conditions. Important physiological traits have been impacted by CNV in such a way that they contribute to crop domestication, or are selected for in crop improvement. CNVs appear to occur more frequently than other genetics changes such as single nucleotide polymorphisms, insertion-deletions or whole genome duplications. Assessing the mechanisms underlying CNV formation is instrumental to understand how this drives plant evolution. Detecting CNVs in plant genomes is not trivial, and newly developed genome sequencing technologies and high-throughput validation methods are needed to aid the detection and interpretation of CNVs in plant genomes. Verification of the assumed high frequency of occurrence will add to its importance as a source of genetic variation that can be selected for to enhance the generation of novel elite cultivars.

Copy number variation, an underestimated source of genetic variation

Over the past decades, the genome sequences of many species have become available and further advances in next generation sequencing now facilitate large-scale resequencing of thousands of individuals of the same species (Alonso-Blanco *et al*., 2016; Huang *et al*., 2012; Zhou *et al*., 2015). Comparative genome analysis initially focussed on the identification of single nucleotide polymorphisms (SNPs) or very short insertions or deletions (indels). With the increased ease of generating (re)sequenced genomes at high coverage, it becomes more and more evident that copy number variation (CNV) is a prominent additional class of intra-species genetic variation. The importance of CNV for plant evolution has been overshadowed by that of another type of structural genome sequence information, resulting from whole genome duplications (WGD). The consequences of WGD for evolution are tremendous (Van de Peer *et al*., 2017). WGD events are often found at branch points of plant phylogenetic trees and are thus likely to be causal for the emergence of new species or lineages. Their co-occurrence at the boundaries between geological periods suggests they are the evolutionary consequence of mass extinction events (Vanneste *et al*., 2014), providing a genetic reservoir of gene redundancy needed for rapid adaptation to the changed environmental conditions. Despite their importance for evolution, WGD events are relatively rare, e.g. in the evolutionary history of the Brassicaceae lineage leading to *Arabidopsis thaliana*, only two of such events have occurred (Van de Peer *et al*., 2009). In this review we will highlight another type of structural genome variation (SV), CNV, which is much more common than WGD, but much less studied. We will provide an overview of the nature of copy number variation, the ways to identify the occurrence of CNV in next generation sequence data, which is not trivial, and we will finally discuss the implications of CNV for the evolution of plants, especially in the context of rapid adaptation to a changing environment.

Structural genomic variation (SV) in general is considered to include all qualitative genomic differences between individuals, such as inversions, translocations, duplications and deletions. CNV acts on the sub-genome level. It is a quantitative type of SV that includes insertions, duplications and deletions and is usually defined as the occurrence of different numbers of a certain DNA segment in at least two individuals of the same species (Scherer *et al*., 2007; Żmieńko *et al*., 2014). The term CNV can also be used when comparing the genomes of closely related species, e.g. the occurrence of one copy of a gene in one species and two or more copies of the functional orthologue in the related species. Next to CNV, the term “presence or absence variation” (PAV) is often used, referring to sequences that are either present in a single copy in one genotype and completely absent in another. To avoid confusing additional terminology to indicate when a sequence is absent or present in one or more copies, we consider PAV to be an extreme form of CNV. In the initial operational definition by (Scherer *et al*., 2007), the size range of a CNV was set from 1 kilo base pairs (kbp) up to submicroscopic size (sometimes millions of base pairs (bp)). It is more practical, however, to set the lower limit at 50 bp (Zarrei *et al*., 2015), classifying smaller segments as indels, and to include an arbitrary higher limit at 100 kbp. CNV thus can span entire clusters of genes, genes may be partially or entirely deleted or duplicated, they may move to new positions or acquire novel regulatory elements. Consequently, both changes in the number of functional gene copies, as well as differences in regulatory sequences due to CNV, can have important effects on gene expression and thereby affect phenotypes.

Copy number variation as a driver of plant evolution

In the past decade, the contribution of CNV to phenotypic variation has become increasingly clear. The initial interest in this type of polymorphisms came from human genetic studies, in which CNV has been associated with a number of disease phenotypes (McCarroll & Altshuler, 2007). Human genetics also showed that CNVs contribute more to genetic variation than single nucleotide polymorphisms (SNPs) in terms of the fraction of the genome affected (reviewed by Campbell & Eichler, (2013a)). Despite this, even for a well-studied species as human, the phenotypic consequences of CNVs are still hardly explored. In plants, the occurrence of CNV is quickly gaining attention, particularly in relation to crop domestication (Lye & Purugganan, 2019). Numerous examples of CNV underlying phenotypic traits in different plant species have been reported, many involved in processes related to environmental adaptation and evolution, as will be discussed below.

The most obvious and direct consequence of a change in relative gene copy number is a gene dosage effect, which contributes to a change in the level of mRNA transcripts corresponding to this copy relative to the transcription of the rest of the genes (Ohno, 1970). Although the relation between transcript level and protein level is not always as strong as one would expect, a change in transcript level is likely to also affect the protein level. Such relative gene dosage effects may thereby cause imbalances, especially in case of regulatory genes such as transcription factors or genes encoding subunits in protein complexes (Papp *et al*., 2003; Tasdighian *et al*., 2017). Interestingly, these types of genes are preferentially retained after whole genome duplication but are rarely found as *de novo* small scale duplications (Tasdighian *et al*., 2017) which indicates clear evolutionary differences (Defoort *et al*., 2019).

While the regulatory imbalances resulting from *de novo* CNVs are initially likely to confer negative effects on fitness, or a neutral effect at best in case of simple genetic redundancy, in the longer run, they provide an important class of genetic variation that can be selected for by providing the opportunity to evolve genes with new functions. Arguably, one of the copies will be under relaxed selection if at least one other functional copy remains active and there are no selective advantages or constraints resulting from a gene dosage effect. The most likely outcome for such duplicate copies is non-functionalization, thus creation of a pseudogene, and eventually loss of one copy. There are two alternative outcomes that offer new evolutionary opportunities. One of which is that the redundant copy may adapt a new function, a process which is generally referred to as neofunctionalization (Ohno, 1970). There are several examples of this, for instance, after the ancient duplication of a *KNOX* transcription factor gene in a common ancestor of land plants, neofunctionalization led to antagonistic rather than redundant roles for the two resulting *KNOX* genes (Furumizu *et al*., 2015). The other option is that after duplication, both gene copies divide the ancestral function amongst them (subfunctionalization). The regulatory sequences and/or the protein functions of the original gene can be partitioned over the duplicate copies in such a way that the gene's ancestral expression pattern and function is covered by the duplicate copies (Freeling *et al*., 2015). For example, subfunctionalization, after a tandem gene triplication of *LATE MERISTEM IDENTITY 1* (*LMI1*)-like sequences, plays a key role in determining the morphological differences in leaf shape between *A. thaliana*, with simple leaves, and the closely related *Cardamine hirsuta*, with dissected leaves. *C. hirsuta* has two functional copies: *REDUCED COMPLEXITY* (*RCO*) and *LMI1*, whereas *A. thaliana* has only retained (Vlad *et al*., 2014). *RCO* and *LMI1* have near-complementary expression patterns in *C. hirsuta*, and overexpression of *LMI1* in *A. thaliana* does not increase leaf shape complexity. Since the *RCO* and *LMI1* protein functions are equivalent, regulatory diversification followed by gene duplication and subsequent gene loss in *A. thaliana* underlies the difference in leaf shape morphology between both species (Vlad *et al*., 2014). Subfunctionalization of duplicated gene copies may evolve further to diverge the ancestral function in such a way that it is no longer covered. This process encompasses subfunctionalization followed by neofunctionalization, and is referred to as specialization or subneofunctionalization (X. He & Zhang, 2005). Analysis of young (less than 5-10 million years ago), retained gene duplicates in *A. thaliana* and *A. lyrata* showed that their most common evolutionary fates are conservation (in which both copies retain their ancestral function), followed by specialization and neofunctionalization, with very few examples of subfunctionalization (Wang *et al*., 2016).

Each of these evolutionary fates of duplicated genes are observed in the MADS-box transcription factor gene family (reviewed by (Airoldi & Davies, 2012)). MADS-box genes are regulators of the determination of meristems and organ identity in developing flowers (Purugganan *et al*., 1995). This multi-gene family has undergone extensive gene duplication events, resulting in over a hundred members in *A. thaliana* (Martínez-Castilla & Alvarez-Buylla, 2003; Parenicova *et al*., 2003). Many gene members originated from ancient or recent whole genome duplications (WGD) although small-scale tandem and dispersed duplications have played an important role in this gene family expansion as well (Arora *et al*., 2007; Nam *et al*., 2004). Such smaller scale duplications often experienced different evolutionary fates, with higher rates of non-synonymous mutations than those duplicated by WGD, as they resulted in unbalanced polymorphisms that may have caused considerable gene dosage effects (Carretero-Paulet & Fares, 2012; Edger & Pires, 2009).

Despite numerous examples of CNV-related phenotypes (reviewed by Dolatabadian *et al*., (2017)), CNV has received little attention compared to single nucleotide polymorphisms (SNPs) and indels. This is undoubtedly due to the larger technical complexity of detecting CNV, especially when it involves nearly identical tandem gene duplications. In this review, we first discuss the different molecular mechanisms that can lead to CNV and provide an overview of the state-of-the-art in detection and validation of CNVs in next generation sequencing data, with the advantages and disadvantages of different techniques. Next, we explore the contribution of CNV to plant adaptation and evolution and provide insights into the frequency and distribution of CNV across species. Finally, we argue that CNV is a common form of structural variation in plants, more common than previously thought, which can be exploited for novel applications in plant breeding.

Generation of copy number variation

Unravelling the mechanistic processes underlying CNV formation in plants is important to understand how this can drive plant evolution. In this respect it is important to distinguish gene duplication due to CNV from those caused by WGD events followed by subsequent gene, which can be substantial (Qiao *et al*., 2019). Most information on CNV generating mechanisms has been obtained through studies in the model organisms *Escherichia coli*, *Saccharomyces cerevisiae* and *Drosophila melanogaster*. Although the CNV mechanisms function similarly in plants, the relative contribution of a particular mechanism may differ between species. CNV initially arises either as a germline or a somatic mutation. While germline CNVs have equal implications in terms of stable transmission in plants and animals, for somatic mutations these implications are different. Where in animals the germline and soma are strictly separated, referred to as the Weismann barrier, this is not the case in plants, in which the germline is formed from somatic cells. Thus somatic mutations in plants may be transmitted both sexually and asexually. Moreover, many plant species, including several agronomically important species, rely partially or entirely on asexual propagation. The relative impact of molecular mechanisms that are more prominent during mitotic cell divisions may therefore be larger in plants than in animals, and are also likely to differ among plant species. The relative contribution of different mechanisms is best studied if the exact breakpoints, size and position of CNVs are determined, which requires detailed genome information. The common factor in the generation of CNV is recombination.

Non-allelic homologous recombination

The most common mechanism underlying CNVs in human genomes is non-allelic homologous recombination (NAHR) occurring during meiosis or during DNA repair after a double strand break (DSB) (Gu *et al*., 2008). Illegitimate recombination of highly similar, but non-allelic, sequences may result in a rearrangement, deletion or duplication and reciprocal deletion depending on the orientation of the homologous sequences and on whether NAHR takes place between different or the same chromatids or between different chromosomes (Figure 1). Also in plants NAHR seems to be a dominant cause of CNV, though the evidence supporting this is limited. Zmienko *et al*., (2016) examined a genomic region in *A. thaliana*, spanning three genes (*MSH2*, *At3g18530* and *At3g18535*), that showed frequent deletions and duplications. Analysis of the sequences around *At3g18530* and *At3g18535* revealed low copy repeats of around 1 kbp, with 99% identical sequences, directly flanking these two genes on either side. The breakpoints of many of the duplication and deletion events co-localize in the LCR region, providing further evidence for an NAHR-based mechanism. Although extensive CNV for all three genes residing at this locus was observed, there was no correlation between copy number and geographic origin of the accession, and CNV at this locus seemed to have occurred multiple times independently from each other, across the examined *A. thaliana* distribution range. This is a strong suggestion that regions with LCRs are more prone to induce the occurrence of recurrent CNVs.

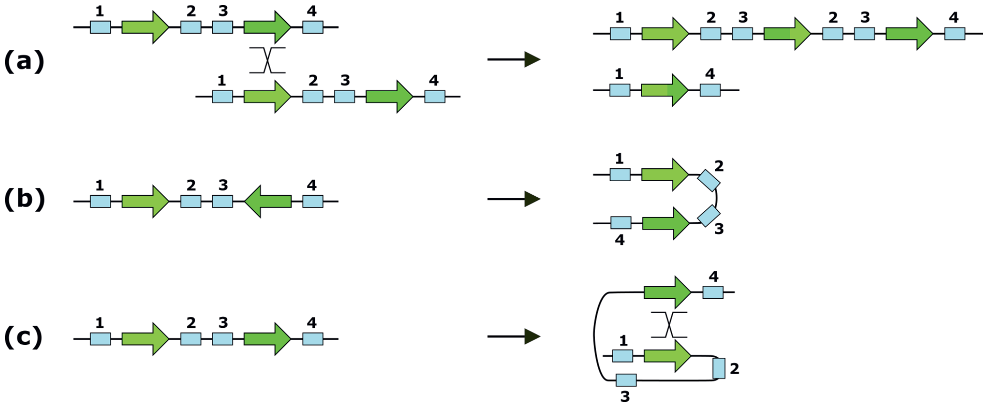


Figure 1: **Non-allelic homologous recombination (NAHR) may result in different types of structural variation.** Illegitimate inter-chromatidal recombination between two low-copy repeats with high sequence homology (differently shaded green arrows) results in a duplication and reciprocal deletion of genes (light blue boxes) 2 and 3 (a). Intra-chromatidal NAHR may result in the deletion of genes 2 and 3, which are lost from the genome (b), or alternatively in an inversion (c) depending on the relative orientations of the low copy repeats.

DNA repair and replication

Several molecular mechanisms related to DNA replication and DSB repair have been postulated to result in SV (Hastings *et al*., 2009; Zhang *et al*., 2009). These mechanisms rely on homologous recombination (HR) or non-homologous recombination (NHR). In plants, DSBs are most commonly repaired by non-homologous end joining (NHEJ): a process that re-joins broken DNA ends and mostly results in accurate DNA repair. However, at low frequencies, NHEJ results in small insertions, deletions or even larger insertions of ‘free’ DNA such as retrotransposons (Yu & Gabriel, 2003). When during DNA replication, a replication fork encounters a nick in the DNA, it may stall or collapse. This induces break-induced repair (BIR), which uses a HR-based mechanism to repair a collapsed replication fork. In the process of BIR, the 5′ end of the broken DNA molecule is resected, leaving a 3′ tail. This tail can invade a homologous sequence, usually the sister chromatid, to restore the replication fork. When instead a non-allelic homologous sequence is used for invasion, different types of SV may arise, depending on the location and orientation of the non-allelic homologous sequence (Figure 2 (Carvalho & Lupski, 2016). A collapsed replication fork can also be restored based on only a few bp of homology. In this case, a microhomology-mediated BIR (MMBIR) mechanism applies to resume replication (Hastings *et al*., 2009). When template switches occur between different replication forks, this known as fork stalling and template switching (FoSTeS) (Zhang *et al*., 2009). Complex SV, such as duplications, deletions, inversions and combinations of these, may result from these processes, depending on the positions of the (micro)homologous sequences used while switching templates (reviewed by (Carvalho & Lupski, 2016; Hastings *et al*., 2009).

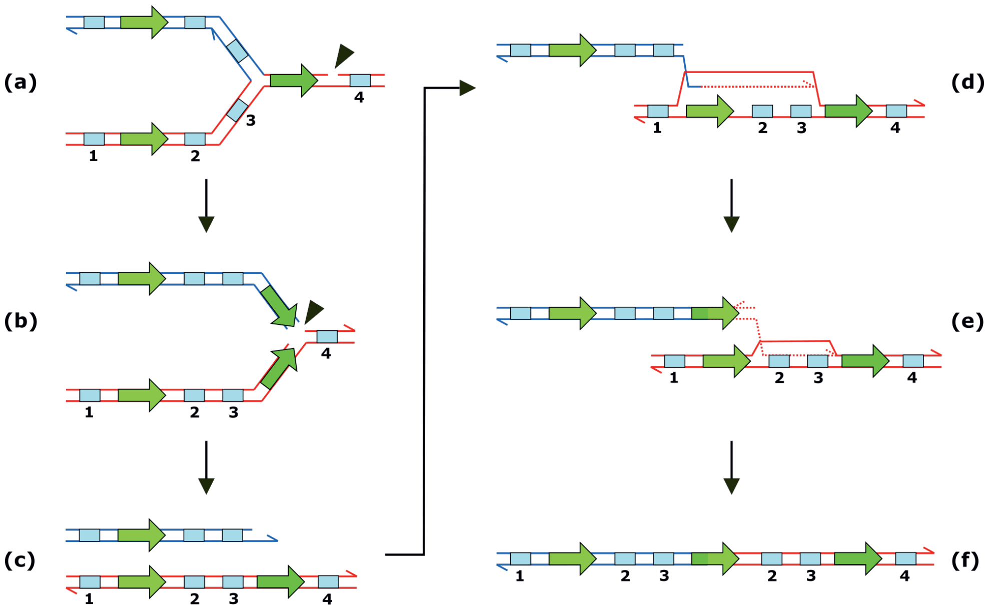


Figure 2: **Break-induced repair model for copy number variation formation.** When a replication fork encounters a nick (black triangle) (a), it may collapse (b). Resection of the 5′ broken DNA molecule leaves a 3′ single strand tail (c). 3′ tail invasion is based on sequence homology (depicted as differently shaded green arrows). However, when a non-allelic homologous sequence is used for 3′ tail invasion (d), the new replication fork is incorrectly positioned (e), which results in this example in a duplication of genes 2 and 3 (light blue boxes) (f).

Transposable elements

Transposable elements (TEs) are often highly represented in regions of the genome that show CNV (Cardone *et al*., 2016; Golicz *et al*., 2016; Hardigan *et al*., 2016; Morgante *et al*., 2007). Obviously, TEs can directly induce their own CNV by self-replication. However, TEs may also directly or indirectly lead to CNV of other, non-TE sequences. Indirectly, they may serve as a substrate for illegitimate recombination mechanisms such as NAHR (Startek *et al*., 2015). Furthermore, certain classes of TEs, such as retrotransposons, helitrons and *Mutator*-like elements (MULEs), are capable of duplicating gene fragments or even entire genes (Brunner *et al*., 2005; N. Jiang *et al*., 2004; Juretic *et al*., 2005; Lai *et al*., 2005). In grasses, the activity of long terminal repeat (LTR) retrotransposons has caused extensive genomic variation in this manner (Morgante *et al*., 2007). Gene capture and gene duplication by TEs mainly produces pseudogenes, although new, chimeric genes, which consist of parts of different genes, may also be formed. For instance, in rice, a type of non-autonomous MULE, called Pack-MULE, frequently contains gene fragments that originate from a single gene or from chimeric genes. Many of these genes containing Pack-MULEs are still transcribed and those that contain chimeric gene fragments are more frequently expressed compared to those containing fragments originating from a single gene (Hanada *et al*., 2009). Such chimeric genes often bear signs of purifying selection, based on the ratio of nonsynonymous and synonymous substitution rates, which means they can be functional (Hanada *et al*., 2009). This process of gene fragment duplication can create new functional proteins by fusion of inserted genic sequences to existing transcribed sequences.

Translocation and recombination

Translocations, i.e. genomic rearrangements in which genomic DNA fragments move to a new location without duplication or deletion, can rapidly create CNV after hybridisation. Various mechanisms can mediate translocations, such as NAHR, DNA repair through NHEJ or TE transposition. The effect of translocations on creating CNV between siblings was best observed in a so-called tetrad analysis, examining all four products of one meiosis. In *A. thaliana*, such tetrad analysis makes use of the *quartet1* (*qrt1*) mutation, in which all four pollen resulting from one microspore mother cell meiosis remain attached to each other as a tetrad (Preuss *et al*., 1994). (Lu *et al*., 2012) performed a tetrad analysis in *A. thaliana* based on a cross between the four attached pollen of a Col-0/Ler *qrt1* hybrid and a wild type Col-0. The progenies of two such crosses were analysed by whole-genome sequencing, and no less than 21 and 32 CNVs of several 100 bps were detected among the progeny due to recombination and assortment of small translocations present as highly similar sequences at different locations in both genomes. After translocation of a sequence to another position on the same chromosome, a cross-over during meiosis is sufficient to generate the CNV (Figure 3). If the sequence is translocated to another chromosome, random chromosome assortment during meiosis will create the CNV (Lu *et al*., 2012).

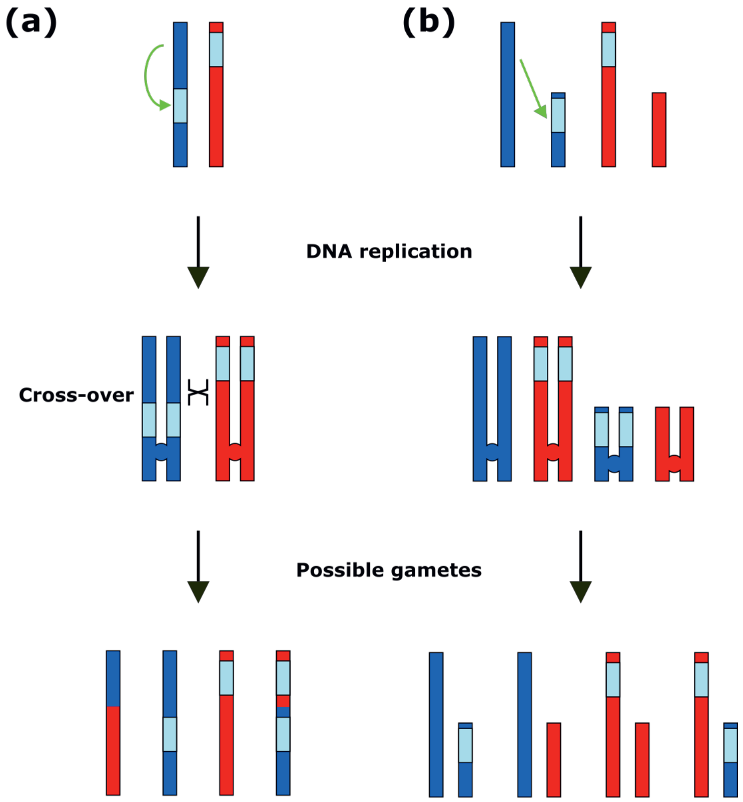


Figure 3: **Translocation and subsequent recombination or chromosome assortment causes CNV.** (a) After translocation (green arrow) of one genomic region (light blue box) to another location on the same chromosome (depicted as dark blue or red bars), a cross-over during meiosis can result in gametes with either none, one or two copies of the genomic region. (b) Similarly, after translocation to another chromosome, only random assortment of chromosomes is sufficient to cause copy number variation.

The relative contribution of each of the above-mentioned CNV-generating mechanisms in plants remains to be resolved. Differences in genome architecture, such as genome size and TE content, will lead to differences in the relevance of the CNV-generating mechanisms between different species, e.g. the contribution of NAHR is very low in cucumber, which is likely due to the low abundance of dispersed repeats in its genome (Zhang *et al*., 2015). The size and new location of a structural variant depends on the underlying CNV-generating mechanism. NAHR is likely to result in tandem duplications of contiguous copies. Upon duplication, most affected genes remain in a similar genomic context, surrounded by the same regulatory elements. Consequently, the main short-term effect of tandem duplications is an increase in gene dosage, often leading to higher expression. NAHR may also lead to segmental duplications, which are large, spanning from 1 to 400 kbp in length, and may translocate blocks of genes to an unlinked site. These do not necessarily alter gene expression or gene function, as their genomic context will remain largely unaltered. Upon smaller, dispersed or "ectopic" duplications, spanning only one or a few genes at unlinked loci and most likely to be caused by transposon capture, genes may lose their original genomic context and acquire new regulatory elements, thus altering gene expression and function.

Improving the detection of copy number variants through novel computational and experimental methods

Over the years a variety of methods is used to detect CNV in plants. Initially microarrays of oligonucleotide probes were widely used for array-based comparative genomic hybridization or aCGH (Beló *et al*., 2010; DeBolt, 2010; Muñoz-Amatriaín *et al*., 2013; Springer *et al*., 2009; Swanson-Wagner *et al*., 2010). This was very successful, especially to detect duplications of nearly identical sequences, but its resolution relied on the number of probes available on the array. Recently CNV detection methods shifted to whole-genome sequencing (WGS) (Cao *et al*., 2011; Chia *et al*., 2012; Hardigan *et al*., 2016). Detecting CNVs through WGS-based methods requires aligning reads of the target genomes, typically short reads produced by an Illumina sequencing platform, to a well-validated reference genome. Algorithms can make use of three types of signals to discover CNVs: discordantly aligned read pairs, split-read alignments and read depth (Figure 4) (Alkan *et al*., 2011). The most effective way to detect CNVs is to combine multiple signals (Lin *et al*., 2015), as done by several recently developed algorithms (Iakovishina *et al*., 2016; Layer *et al*., 2014; Rausch *et al*., 2012), aiming at high sensitivity and high precision.

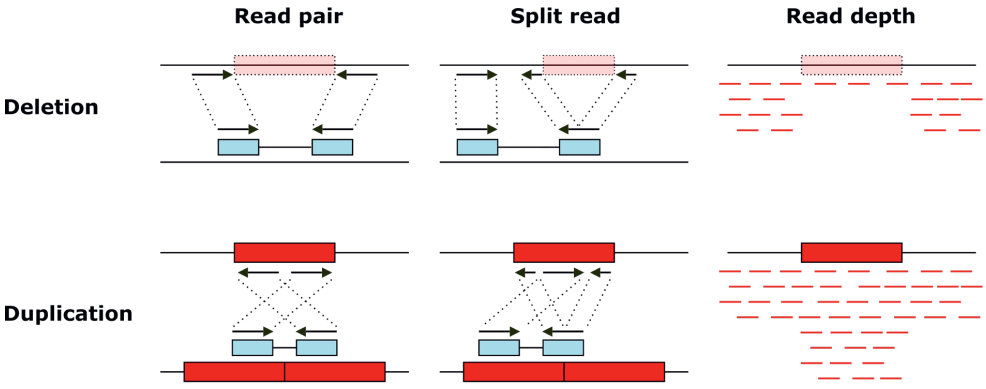


Figure 4: **Overview of the signals used to detect different types of copy number variants.** Deletions are coloured light-red and duplicated regions are coloured dark red. The read pair and split read columns contain alignments of paired-end reads to the reference (top) and target genome (bottom) with reads shown in light-blue and their orientation depicted by arrows. The red lines in the read depth column depict reads mapped to the reference genome.

Short-read sequencing has enabled the extensive study of CNVs in a wide variety of plant species, as seen by the many examples given in this review. Yet, it has important limitations, including poor detection of CNVs in repetitive regions, low sensitivity for CNVs in polyploid genomes, and limited ability to detect CNVs between genomes that differ significantly at the nucleotide level. Hence, many CNVs present in plant genomes may remain hidden (Huddleston & Eichler, 2016). Recent advancements in long-read sequencing technologies address the technical limitations associated with short reads (Goodwin *et al*., 2016). The increased sequence length (ranging from 10 kbp to over 10 Mbp) particularly helps to detect CNVs within repetitive elements and segmental duplications. Eventually, the combination of long-read sequencing and long-range scaffolding technologies, such as optical mapping, Hi-C and 10x Genomics linked-reads, will facilitate full-chromosome genome assemblies of genomically complex plant species (W. B. Jiao & Schneeberger, 2017). Such assemblies will contribute to a better detection of CNVs without the need to rely on one (potentially poor) reference genome. The costs of long-read sequencing to characterize CNVs in large population studies is still prohibitive, but will undoubtedly come down soon. A solution to overcome the limitations of short-read technology is to exclude repetitive regions from CNV analysis by genome reduction methods such as the genotyping-by-sequencing protocol (Elshire *et al*., 2011), or to use computational methods that can combine both short-read, and long-read technologies (Fan *et al*., 2017).

While the benefits of using long-read sequencing technologies are evident, such technologies will only improve the detection of CNVs in plant species with the development of novel computational algorithms that are specifically tailored for plants. New algorithmic ideas and data structures are also needed to assemble, phase, and represent polyploid genomes, for which there are virtually no dedicated tools available (Sedlazeck, Lee, *et al*., 2018). Such novel algorithms should be improved regarding their sensitivity for allele-specific CNVs, that are present in only a subset of the genomic haplotypes. With comparatively low sequence coverage of such CNVs compared to CNVs present in all haplotypes, current computational methods often fail to detect them. Detection of such allele-specific CNVs is important as they can lead to markedly affected haplotypes (VanBuren *et al*., 2018).

In view of the stochasticity in detecting CNVs, experimental validation is recommended to confirm their presence. Real-time quantitative PCR (qPCR) is frequently used, although it has some limitations. First of all, several to many replicates are required to accurately quantify small copy number differences between the target and the reference genome. Especially when several (> 3) copies are present, increasingly more replicates are needed to detect single copy differences (Weaver *et al*., 2010). Secondly, qPCR tends to underestimate the true copy number of highly duplicated regions, due to differences in kinetics at the early and late stages of PCR amplification (Lee & Jeon, 2009). Finally, setting up a successful qPCR assay is not straightforward: some steps (e.g. PCR primer design) require extensive optimization to prevent amplification of non-target regions. Two recently developed techniques, multiplex ligation-dependent probe amplification (MLPA) and droplet digital PCR (ddPCR), offer interesting alternatives: a case study using *A. thaliana* datasets revealed that both can validate CNV regions with copy numbers up to 8 (Zmienko *et al*., 2016).

Biological implications of copy number variation

Gene ontology enrichment analyses in different species, of genes residing in CNVs, show remarkably comparable results with respect to the biological processes in which these genes are involved. Asides from TEs, which are a common class of genes subject to CNV, genes involved in stress-response pathways and especially disease resistance are overrepresented in *A. thaliana* (Cao *et al*., 2011), rice (Bai *et al*., 2016), common bean (Ariani *et al*., 2016), domesticated apple (Boocock *et al*., 2015), maize (Chia *et al*., 2012), barley (Muñoz-Amatriaín *et al*., 2013), potato (Hardigan *et al*., 2016), cucumber (Zhang *et al*., 2015) and grapevine (Cardone *et al*., 2016). Plant disease resistance (*R*) genes often act in a so-called 'gene-for-gene' interaction, in which one *R* gene controls the disease resistance response upon expression of an avirulence gene by the pathogen. Pathogens typically have a high evolvability, which means that mutations in the avirulence gene can quickly render the corresponding *R* gene ineffective. Plants, in turn, harbour many *R* genes, often encoding receptor-like proteins with leucine-rich repeat (LRR) domains, which are also highly polymorphic, and often reside in clusters of paralogous gene copies, which typically consist of highly similar and repetitive sequences. Such genomic architecture is prone to mutagenic processes such as polymerase slippage and NAHR (reviewed by (Karasov *et al*., 2014). Rapid gene duplication followed by mutation is thus an important driver of novel *R* gene variation, that allows plants to co-evolve with their pathogens.

Copy number variation underlies plant adaptation to environmental stress

Being unable to evade environmental disturbances, plants require quick and effective mechanisms for adaptation to their environment. Copy number expansion (CNE) of genes implicated in the response to environmental fluctuations may offer an effective way to improve plant fitness within a few generations depending on the selection pressure. Application of a herbicide such as glyphosate provides a strong selection pressure, adaptation to which was achieved by means of CNE in *Amaranthus palmeri* (Gaines *et al*., 2010, 2011). The action of glyphosate lies in its binding to the enzyme 5-enolypyruvylshikimate-3-phosphate synthase (EPSPS; EC 2.5.1.19). This enzyme in the shikimate pathway, breaks down phosphoenolpyruvate (PEP) into aromatic amino acids (phenylalanine, tyrosine and tryptophan) which are required for plant growth. Binding of glyphosate to EPSPS forms a dead-end complex, inhibiting the shikimate pathway, and resulting in plant death (Steinrücken & Amrhein, 1980). Originally, the glyphosate susceptible *A. palmeri* only had a single copy of the EPSPS gene, but strong selection pressure by herbicide application has resulted in resistant plants which harbour from 5 to as many as 160 copies of the gene due to CNE (Gaines *et al*., 2010). This increased the expression of *EPSPS* and saturation of the glyphosate pool, effectively allowing the excess of remaining unbound EPSPS to perform its function as normal. The *EPSPS* copy number and the abundance of functional EPSPS correlate very well with the level of herbicide tolerance, exemplifying the gene dosage effect of CNE. CNE of *EPSPS* in *A. palmeri* must have occurred very recently, after the introduction and large-scale application of glyphosate, explaining the extremely high sequence similarity of the *EPSPS* gene copies, with even introns showing no DNA polymorphisms (Gaines *et al*., 2013). Resistance to glyphosate can be conferred by SNPs in the EPSPS protein coding sequence, though reports of this are rare and the conferred resistance is not as strong as that achieved by CNE (Sammons & Gaines, 2014). CNE-mediated glyphosate resistance has since been reported for several other weed species, although the underlying molecular mechanisms of CNE are not the same (Dillon *et al*., 2017; Gaines *et al*., 2016; Jugulam *et al*., 2014; Molin *et al*., 2017; Sammons & Gaines, 2014). The *A. palmeri EPSPS* gene copies are distributed across all chromosomes, with a possible role in translocation for helitron TEs (Gaines *et al*., 2010; Molin *et al*., 2017) while in, e.g., *Kochia scoparia* copies are generally found to be arranged in tandem duplications , suggestive of a NAHR-mediated mechanism (Jugulam *et al*., 2014).

Similar examples of adaption can be found in extremophiles: plant species adapted to highly challenging environments (Oh *et al*., 2013). Genome comparisons between extremophiles and closely related non-extremophiles often indicate adaptation-related CNE. *Thellungiella parvula* and *T. salsuginea*, two related Brassicaceae species adapted to saline, resource-poor habitats, showed similar genes to be duplicated in comparison to the salt-sensitive *A. thaliana* (Dassanayake *et al*., 2011; H.-J. Wu *et al*., 2012). These mostly tandem duplications were preferentially involved in known stress defence responses such as ionic stress protection (Dassanayake *et al*., 2011; H.-J. Wu *et al*., 2012). For instance, CNE of *HKT*, a locus encoding for a high affinity Na+/K+ transporter associated with salinity tolerance in *A. thaliana* (Baxter *et al*., 2010), has occurred in both species: a duplication in *T. parvula* and a triplication in *T. salsuginea* (Dassanayake *et al*., 2011; H.-J. Wu *et al*., 2012).

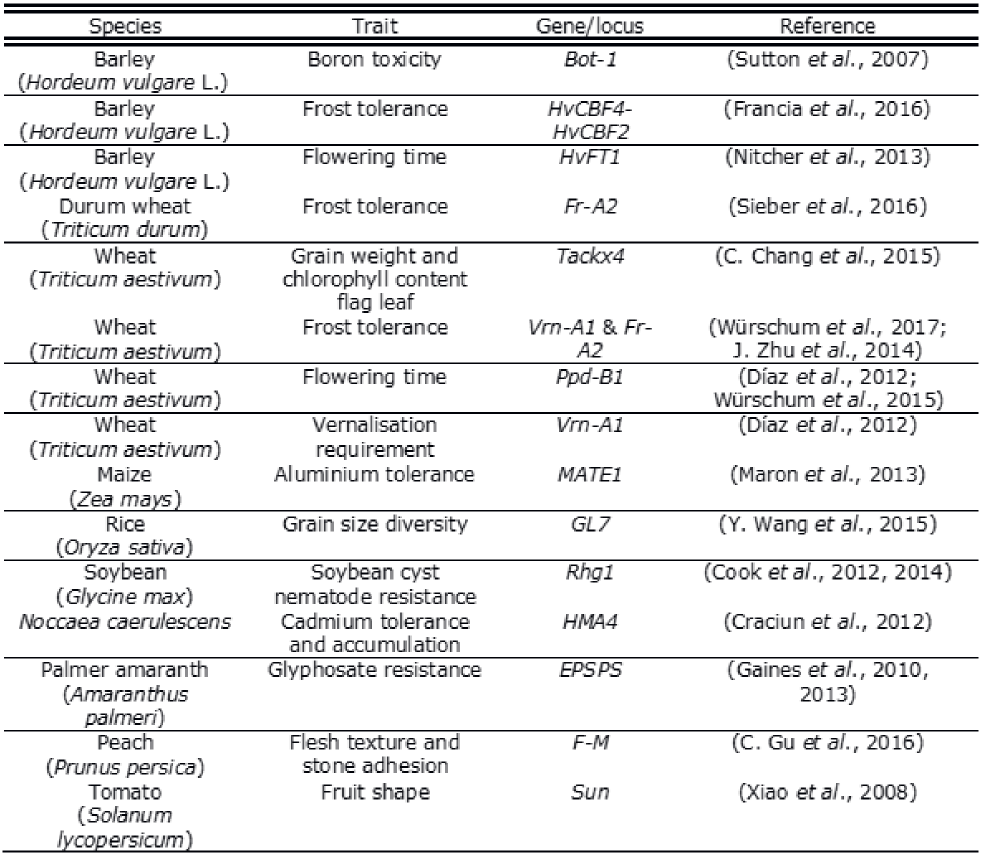
Adaptive CNEs are also observed in the heavy metal hyper-accumulating extremophiles *Arabidopsis halleri* and *Noccaea caerulescens*. These species are able to grow on soils containing excessive, and normally toxic, concentrations of heavy metals such as cadmium and zinc, and may accumulate these metals to 50-100x higher concentrations than surrounding non-hyperaccumulating vegetation (Cappa & Pilon-Smits, 2014). Metal homeostasis-related genes were significantly enriched among duplications in *A. halleri* compared to closely related non-hyperaccumulators *A. thaliana* and *A. lyrata* (Suryawanshi *et al*., 2016). The importance of gene duplication in adaptation to heavy metal contaminated soils is illustrated by CNE of the *HEAVY METAL ATPASE4* (*HMA4*) gene that functions in root-to-shoot Zn/Cd translocation. In *A. halleri*, three tandem copies of *HMA4* are present (Hanikenne *et al*., 2008), while in *N. caerulescens* two to four tandem copies can be found (Craciun *et al*., 2012). This CNV for *HMA4* in *N. caerulescens* was observed among three different populations with variable levels of cadmium accumulation and tolerance. The accession with the lowest number of *HMA4* genes also had the lowest expression level of this gene and the least cadmium translocation.

It is important to understand how quickly populations can adapt to new environments. One major factor affecting the pace of adaption is the frequency of CNV events occurring *de novo*. DeBolt, (2010) studied the effect of different temperatures (16 and 28 °C) and mimicked biotic stress (salicylic acid spray) compared to normal growth conditions (22 °C, mock assay) on the occurrence of CNV in *A. thaliana*. Starting from one parent, plants were exposed to the different stress treatments for five consecutive generations. In each generation, the plant with highest fecundity was selected for the next generation. Finally, the genomes of three siblings per stress treatment were compared to those of three siblings from the reference lineage by aCGH, to detect repetitive and non-repetitive CNV between siblings at the end of the fifth generation. The lower temperature and salicylic acid treatments resulted in 14, resp. 13 CNV events, spanning ~400 genes in total in both cases. All of these were deletions. The high temperature treatment resulted in 11 CNVs, of which seven were duplication events and several were unique among siblings. This illustrates the high prevalence of CNV, even within a very small time frame, suggesting very rapid evolution. The high temperature treatment imposed most stress on the plants (determined based on effects on biomass, germination rate and seed set) and was also the treatment with most duplications and most unique CNVs among siblings. This raises the question if there may be a causal relationship between the level of stress imposed and the frequency of CNV events. This is not unrealistic: higher levels of stress will increasingly induce the formation of reactive oxygen species, which can lead to more DNA damage and thus more errors in repair (Sharma *et al*., 2012). In addition, for several stresses the frequency of homologous recombination will be induced (reviewed by Migicovsky & Kovalchuk, (2013a)) and stress exposure is known to activate transposons through genome demethylation (Hashida *et al*., 2006; Madlung & Comai, 2004). All of these processes can contribute to CNV. In case of strong stresses, this will impose a strong selection pressure so that beneficial CNV events can be selected, resulting in the amplification of genes moderating the effect of the stress and the loss of genes enhancing stress sensitivity. As this seems to happen already within a few generations, far quicker than expected based on gradual evolution of stress adaptation through SNPs, this appears to be a major driver of the evolution of environmental stress tolerance in plants.

Copy number variation contributes to phenotypic variation

CNVs frequently affect phenotypic traits which can have important ecological or agronomical implications (Table 1). Qualitative traits due to CNV will mostly be found in the case of presence or absence variants. For instance, variation in resistance (resistant or susceptible) to *Pseudomonas syringae* in *A. thaliana* is caused by the presence or absence of a resistance gene, *RPS5* (Simonich & Innes, 1995; D. Tian *et al*., 2002). Copy number variation beyond one copy will however be mostly of a quantitative rather than qualitative nature, due to the functional redundancy of additional (nearly) identical copies. For instance, *Heterodera glycines*, the soybean cyst nematode, is one of the most damaging pests of soybean (*Glycine max*) (Cook *et al*., 2014). Different levels of resistance are conferred by a 31.2 kbp genomic region, the *Rhg1* locus, harbouring three genes that contribute to resistance (Cook *et al*., 2012). Susceptible varieties carry one copy of the region at this locus whereas in resistant varieties up to 11 tandemly duplicated copies are present (Cook *et al*., 2012; T. G. Lee *et al*., 2016). Overexpression of all three genes together confers resistance in a susceptible variety, illustrating the quantitative nature of the trait (Cook *et al*., 2012). Interestingly, among the offspring of a single plant of the inbred cultivar Fayette the copy number of *Rgh1* varied between 9 and 11 copies (Lee *et al*., 2016). This demonstrates that the copy number of *Rhg1* is unstable, changing already within a few generations, and that it is possible to select for resistance through selection on higher copy number.

Table 1: **Phenotypic traits affected by CNV.**



Species	Trait	Gene/locus	Reference
Barley (*Hordeum vulgare* L.)	Boron toxicity	*Bot-1*	(Sutton *et al*., 2007)
Barley (*Hordeum vulgare* L.)	Frost tolerance	*HvCBF4-HvCBF2*	(Francia *et al*., 2016)
Barley (*Hordeum vulgare* L.)	Flowering time	*HvFT1*	(Nitcher *et al*., 2013)
Durum wheat (*Triticum durum*)	Frost tolerance	*Fr-A2*	(Sieber *et al*., 2016)
Wheat (*Triticum aestivum*)	Grain weight and chlorophyll content flag leaf	*Tackx4*	(C. Chang *et al*., 2015)
Wheat (*Triticum aestivum*)	Frost tolerance	*Vrn-A1* & *Fr-A2*	(Würschum *et al*., 2017; J. Zhu *et al*., 2014)
Wheat (*Triticum aestivum*)	Flowering time	*Ppd-B1*	(Díaz *et al*., 2012; Würschum *et al*., 2015)
Wheat (*Triticum aestivum*)	Vernalisation requirement	*Vrn-A1*	(Díaz *et al*., 2012)
Maize (*Zea mays*)	Aluminium tolerance	*MATE1*	(Maron *et al*., 2013)
Rice (*Oryza sativa*)	Grain size diversity	*GL7*	(Y. Wang *et al*., 2015)
Soybean (*Glycine max*)	Soybean cyst nematode resistance	*Rhg1*	(Cook *et al*., 2012, 2014)
Noccaea caerulescens	Cadmium tolerance and accumulation	*HMA4*	(Craciun *et al*., 2012)
Palmer amaranth (*Amaranthus palmeri*)	Glyphosate resistance	*EPSPS*	(Gaines *et al*., 2010, 2013)
Peach (*Prunus persica*)	Flesh texture and stone adhesion	*F-M*	(C. Gu *et al*., 2016)
Tomato (*Solanum lycopersicum*)	Fruit shape	*Sun*	(Xiao *et al*., 2008)

Some of the main features of this nematode resistance illustrate key characteristics that return in several quantitative CNV examples: an increase in copy number, that can be selected for, leads to increased gene expression which underlies phenotypic variation (Díaz *et al*., 2012; Maron *et al*., 2013; Nitcher *et al*., 2013; Sutton *et al*., 2007; Wang *et al*., 2015). Similarly, variation in freezing tolerance, an important trait for the adaptation of cereals to temperate regions, is mainly caused by CNV at the *Fr-A2* locus in winter durum wheat (Sieber *et al*., 2016), wheat (Würschum *et al*., 2017; J. Zhu *et al*., 2014) and barley (Francia *et al*., 2016; Knox *et al*., 2010). Increased copy number of *C-repeat binding factor* (*CBF*) MADS-box transcription factors at this locus mediate cold acclimation by controlling multiple effector genes. The increased *CBF* copy number correlates well with increased expression and increased cold tolerance (Stockinger *et al*., 2007).

Copy number variation affects a large part of the genome

Genome comparisons in bacteria were the first to reveal substantial SV between different strains, which led to the introduction of the pan-genome concept (Tettelin *et al*., 2005). The pan-genome describes all the DNA sequences present within a certain phylogenetic clade (Vernikos *et al*., 2015). These genome sequences can be divided into a core genome, containing sequences shared among all individuals of that particular phylogenetic clade, and a dispensable genome, consisting of sequences only present in a subset of the phylogenetic clade. The core genome consists of genes related to the basic maintenance of the organism and the main phenotypic traits, while the dispensable genome includes genes that are not essential for survival in general, but may contribute to diversity and local adaptation (Tettelin *et al*., 2005).

A common problem during the analysis of resequencing data is that newly sequenced plant ecotypes, varieties or individuals harbour sequences not found in the reference genome. Flow cytometry and whole-genome sequencing showed up to 10% variation in genome size in *A. thaliana* (Q. Long *et al*., 2013; Schmuths *et al*., 2004). Also in maize, four randomly chosen DNA segments, ranging from 100 to 350 kbp, were sequenced and compared between the two inbred lines Mo17 and B73. Surprisingly, ~38% of all non-TE genes in these regions were present in one of the inbred lines and absent in the other (Brunner *et al*., 2005). Similar high frequencies of SV were also found on a whole genome scale in maize. Initially, hundreds of genes affected by CNV and several thousands of presence/absence variants were identified with gene-based aCGH when comparing several inbred lines (Beló *et al*., 2010; Lai *et al*., 2010; Springer *et al*., 2009; Swanson-Wagner *et al*., 2010). This is an underestimate though, as the aCGHs only covered the sequences found in the B73 reference genome. Whole genome sequencing of ten maize inbred lines later indeed revealed much higher numbers of SV (Chia *et al*., 2012). As these sequences cannot be aligned to the reference genome, they are often disregarded and simply put aside as repetitive elements or transposons. This underlines the dynamics of genomes and illustrates that the genome sequence of one single reference genotype is insufficient to capture all genomic content of a species. It may mean though that important genetic information underlying phenotypic variation is missed. Given the many CNV events that are found between individuals (reviewed by Marroni *et al*., (2014)) it seems appropriate to also apply the concept of a pan-genome to plants (Table 2).

Table 2: **Genome-wide CNV analysis studies in plants.**

Species	Number of genotypes used in the study	Number of genes affected by CNV	CNV detection method	Reference
Common bean (*Phaseolus vulgaris*)	18 genotypes (wild and domesticated)	343 genes	Genotyping-by-sequencing	(Ariani *et al*., 2016)
Domesticated apple (*Malus x domestica* Borkh)	30 accessions	845 genes	NGS	(Boocock *et al*., 2015)
Grapevine (*Vitis vinifera* L.)	4 accessions	2029 genes	NGS + aCGH	(Cardone *et al*., 2016)
Potato (*Solanum tuberosum*)	12 monoploids/double monoploids	30.2% of the genome	NGS	(Hardigan *et al*., 2016)
Barley (*Hordeum vulgare* L.)	14 genotypes (wild and domesticated)	9% of genes	aCGH	(Muñoz-Amatriaín *et al*., 2013)
Rice (*Oryza sativa*)	50 accessions	2806 genes	NGS	(Bai *et al*., 2016)
Brassica oleracea	10 (9 accessions and 1 wild relative)	18.7% of genes	NGS	(Golicz *et al*., 2016)
Brassica rapa	3 accessions	~1000 genes per genotype	NGS	(Lin *et al*., 2014)
Medicago truncatula	15 accessions	49000-169000 CNVs	NGS	(P. Zhou *et al*., 2017)
Glycine soja	7 accessions	1978 genes	NGS	(Y. Li *et al*., 2014)
Soybean (*Glycine max*)	4 accessions	672 CNV 133 PAV	aCGH + targeted resequencing	(McHale *et al*., 2012)
Sorghum (*Sorghum bicolor*)	4 inbred lines	16487 PAV 17111 CNV	NGS	(L.-Y. Zheng *et al*., 2011)
Arabidopsis thaliana	80 accessions	1059 CNVs	NGS	(Cao *et al*., 2011)
Maize (*Zea mays*)	19 inbreds + 14 wild relatives	3889 CNVs	aCGH	(Swanson-Wagner *et al*., 2010)
Maize (*Zea mays*)	2 inbred lines	2714 PAV	Comparison of genome assemblies	(Hirsch *et al*., 2016)

The high prevalence of CNV observed in maize is not common to all plants though. Among 14 barley accessions (wild and domesticated), less, but still substantial, CNV was found, spanning almost 15% of the genome (Muñoz-Amatriaín *et al*., 2013). Among these CNVs, 46.3% was only found in wild barley, 16.8% was unique for cultivated barley, while the remainder was present in both. The higher number of CNVs in wild barley is likely again an underestimate, as the array used for the analysis was designed based on a cultivated accession. In general, CNV will be more frequent in wild accessions than cultivated germplasms, as genetic variation will be lost during domestication. In other species, the number of CNV events can be much lower. For instance, only 343 genes potentially displaying CNV were found in a panel of 18 wild and cultivated common bean (*Phaseolus vulgaris L.*) accessions, based on sequence coverage using a genotyping-by-sequencing approach (Ariani *et al*., 2016). Similarly, in 30 apple (*Malus x domestica* Borkh) varieties, 876 CNVs, roughly 3.5% of the genome, were found by a read-depth-based method (Boocock *et al*., 2015).

The comparison of CNV occurrence between species is non-trivial and ambiguous, as different numbers of genotypes have been studied and different techniques to detect CNV have been employed. Nevertheless, based on a few well-studied cases, there are clear species-specific differences. Striking is the difference in CNV prevalence between maize and *A. thaliana*, likely influenced by different genome architecture and mode of reproduction. The maize genome is estimated to largely consist (85%) of TEs, repetitive sequences which will aid in the formation of CNV formation (Springer *et al*., 2009). *A. thaliana* contains far less TEs, and is highly homozygous because of its propensity to inbreed. This in contrast to cultivated potato (*Solanum tuberosum*), which is highly heterozygous and propagated asexually through tubers, because of which somatic mutations have a high likelihood of becoming accumulated. CNV was assessed by sequencing a panel of 12 related monoploid/doubled monoploid potato clones (Hardigan *et al*., 2016). When comparing the clones, around 30% of the genome was impacted by SV, with similar ratios of duplications and deletions among the different genotypes.

Genome-wide distribution of copy number variation events

The frequency of CNV events is not equally distributed across the genome. In potato, it is highest in the pericentromeric regions, and lowest in gene-dense euchromatic arms (Hardigan *et al*., 2016). Especially large CNVs (larger than 100 kbp) are more prevalent in pericentromeric regions in potato (Hardigan *et al*., 2016). Intuitively this makes sense, since large CNVs in gene-dense euchromatic arms will affect multiple genes and such SV is therefore likely to confer a strong selective disadvantage. Even when a large CNV, spanning multiple genes, has a selective advantage, recombination and selection are likely to break down the CNV into a much smaller region carrying the SV conferring the selective advantage. Closer to the centromere, the recombination rate and the gene density decrease, and larger CNVs are thus more likely to remain in the pericentromeric regions than elsewhere. This was confirmed in maize, which showed that CNVs occurred most frequently in the pericentromeric regions, and that the occurrence was negatively correlated with recombination rate and gene density (Hirsch *et al.*, 2016; F. Lu *et al.*, 2015). In contrast, the CNV frequency in barley was highest towards the chromosome ends, and presence was positively correlated with the recombination rate (Muñoz-Amatriaín *et al.*, 2013). However, this study was performed using aCGH, which favours detection of small CNVs, with about 60% of all CNVs found to be smaller than 200 bp. The discrepancy between CNV frequencies in potato and barley could therefore also be due to the high frequency of small-scale CNVs found in barley. Many CNVs are shared among different genotypes of the same species, also between wild and domesticated varieties and even between closely related species (Bai *et al.*, 2016; Muñoz-Amatriaín *et al.*, 2013; Pinosio *et al.*, 2016). For instance, 1.1% of the SV between the closely related poplar species *Populus nigra* (four genotypes), *P. deltoides* (two genotypes) and *P. trichocarpa* (one genotype) is at least shared between two species (Pinosio *et al.*, 2016). This means such SVs are ancient, and are probably maintained by selection over a long period of time.

Exploring copy number variation for plant breeding

The dynamics of CNV, with rapid change in copy number already within a few generations, and its contribution to favourable traits, opens up the possibility to incorporate the generation of genetic variation due to CNV into plant breeding programs. This means that prior to selection of favourable genotypes, CNV events will be allowed to occur, either without prior knowledge on target loci, or, in a targeted approach, to occur at loci known to be perceptive to CNV, i.e. surrounded by sequences prone to NAHR or TEs. Favourable new variants can then be selected based on molecular analysis or on phenotype. The maintenance and detection of favourable CNVs will be expedited by imposing selective pressure. For example, exposure to the target pathogen may help in selecting the new genotypes carrying new CNV-derived resistance genes. The duration or scale of the generation of genetic variation dsue to CNV is likely to depend on the molecular mechanism underlying CNV generation, which will need to be tested experimentally to come to a practical approach. The obvious downside of selecting for CNV-derived genetic variation contributing to new traits is that continued instability of the locus in the absence of strong selection could lead to genetic heterogeneity and even loss of the trait, which will need to be monitored.

With the increasing awareness of the phenotypic implications of CNV, incorporating the detection of CNV into genetic mapping studies seems a logical next step. Building detailed CNV maps for populations that are used in genome wide association mapping and family F_2, recombinant inbred line (RIL) or near-isogenic line (NIL) mapping could aid in this. The first study in plants that mapped SVs to detect quantitative trait loci (QTLs) in a segregating population was recently published. Imprialou *et al.*, (2017) used ultra-low-coverage (0.3x) population sequencing in an *A. thaliana* Multiparent Advanced Generation Inter-Cross (MAGIC) RIL population, to detect SVs based on read-mapping anomalies. These SVs were mapped to the genome and used as quantitative traits to investigate their association to nine physiological phenotypes. Despite some inevitable limitations posed by the very low coverage, their method detected SVs potentially causal for QTLs affecting germination time, bolting time and resistance to the fungal pathogen *Albugo laibachii*. Alongside genetic mapping, the authors linked transcriptome analysis to the obtained SV, revealing that genes residing within large SV regions are more likely to become silenced or dysregulated.

Conclusion

In this review, we have shown that CNV is a frequent and common class of genetic mutations, found in every plant species. There are several examples of loci displaying CNV associated with variation in phenotypic traits. Whole-genome analyses indicate similar CNV signatures in adaptation-related traits in various wild plant species. Currently, not much is known about the different molecular mechanisms that contribute to the generation and (in)stability of CNV. To gain more insights into the relative contributions of different mechanisms, detailed and systematic surveys employing accurate CNV detection and validation tools are necessary.

Such research will open doors to answer many open questions. It will allow a much better understanding of the generation frequency of CNV occurrence in different plant species. This is especially interesting when studying adaptation to unfavourable environmental conditions, in which the occurrence of CNV at selected loci appears to be a strong evolutionary driving force. Obtaining high resolution CNV maps will aid in the detection of QTLs caused by CNV, as is exemplified in the approach of Imprialou *et al*., (2017) as an interesting first step in this direction. Combining low-coverage, low-cost sequencing with accurate CNV maps will provide the much needed insights into the actual contribution of CNV to important agronomical traits. Dedicated molecular or phenotypic selection procedures may be designed to speed up the detection of *de novo* generated favourable CNV events and develop this as a practical contribution to the improvement of plant varieties by CNV-breeding.

Chapter 3

Rapid adaptation of *Arabidopsis thaliana* to excess zinc stress

René Boesten[1*], Raúl Y. Wijfjes[2,3*], Frank F.M. Becker[1],
Jeroen van der Woude[1], Dirk-Jan M. van Workum[2],
Sandra Smit[2], Dick de Ridder[2,†], and Mark G.M. Aarts[1,†]

[1] Laboratory of Genetics, Wageningen University & Research
[2] Bioinformatics Group, Wageningen University & Research
[3] Current affiliation: Faculty of Biology, Ludwig Maximilian University of Munich

[*]These authors contributed equally to this work.

Supplementary methods and tables are available at:
https://doi.org/10.5281/zenodo.8204499

Abstract

Heritable mutations are a fundamental source to generate adaptive phenotypes. The rate at which heritable de novo (epi)genetic variation arises influences how rapid plants can adapt to adverse environmental conditions. However, if and how mutation rates are affected by stress perceived by plants remains mostly uncertain. Here, we examined *Arabidopsis thaliana* populations grown under moderate and severe salinity and excess zinc stress for five generations using an experimental evolution approach. Based on whole-genome sequencing, we test if these specific stresses affect mutation rates, and observe a twofold increased accumulation of mutations in plants exposed to severe excess zinc stress, but no increases to moderate or severe salinity stress. Additionally, we demonstrate that one of the experimental populations exposed to severe zinc excess stress has significantly improved its performance and fitness after five generations of excess zinc exposure. Adaptation in this excess zinc-tolerant genotype most likely occurred due to a *de novo* mutation causing a premature stop codon in the *RBOHF* gene, encoding an NADPH oxidase, involved in reactive oxygen species (ROS) formation. Our results suggest that *A. thaliana* can rapidly adapt to environmental stress through *de novo* mutations of strong effect.

Introduction

Plants are frequently exposed to unfavourable environmental conditions, that impose stress on the plant, which hinders their development and even proves lethal in extreme cases. Understanding how plants adapt to such stress is a fundamental question in evolutionary biology and plays a crucial role in addressing the increasing demands on crops in the face of climate change (Zaidem *et al*., 2019). The acute responses to stress involve various processes, starting from the initial perception of stress, followed by signalling, signal integratiown, and ultimately resulting in a response achieved through adjustments in transcriptional, translational and biochemical/physiological processes (H. Zhang *et al*., 2022). Additionally, epigenetic changes, such as DNA/histon (de)methylation, may occur that can be inherited for a few generations in the form of "stress memory", but these are typically reverted once the stress subsides (Lämke & Bäurle, 2017). Over multiple generations, beneficial mutations may be acquired through adaptive evolution to better match the environmental conditions. Nevertheless, there is still much unknown about how various stresses influence mutation rates in plant and the significance of these mutations in generating adaptive phenotypes.

This process can be directly investigated using experimental evolution, where populations derived from a single genotype are propagated in a controlled and selective environment (Kawecki *et al*., 2012). Assuming that the chosen environment requires an adaptive response, one can observe which *de novo* beneficial mutations are generated and selected within the evolving populations in real-time. As it may take many, often hundreds of generations before the first beneficial mutations rise to fixation (Lenski *et al*., 1991), most adaptive evolution experiments have been conducted using microbial species. Those experiments have revealed that the genetic architecture of stress adaptation is dependent on the rate at which beneficial mutations accumulate, the fitness effects of such mutations, and the strength of selection (Barrick & Lenski, 2013; Conrad *et al*., 2011).

A limited number of multi-generational growth experiments have been conducted to investigate the genetic mechanisms of stress adaptation in plants. However, practical constraints have limited these experiments to a few tens of generations, in contrast to the thousands of generations typically studied in microbial experiments. Mutation accumulation experiments of *Arabidopsis thaliana* exposed to salinity (C. Jiang *et al*., 2014) and high-temperature stress (Belfield *et al*., 2021; Z. Lu *et al*., 2021) indicate that stress increases the rate of spontaneous accumulation of single-nucleotide polymorphisms (SNPs) and small insertions/deletions (indels). However, these studies did not directly address whether such mutations contribute to adaptation. In addition to genetic mutations, there is evidence that heritable changes in DNA methylation may also contribute to generating a rapid adaptive response (Schmid *et al*., 2018; X. Zheng, Chen, *et al*., 2017). Nevertheless, it remains unclear whether these epimutations are stable or represent transient stress memory, as distinguishing between the two is challenging (Johannes & Schmitz, 2019). Consequently, the relative contribution and speed at which genetic and epigenetic mutations contribute to stress adaptation in plants remain mostly unclear.

One particularly underexplored aspect is the contribution of copy number variation (CNV), defined here as deletions, insertions, and duplications of at least 50 base pairs (bp). CNVs have the potential to make significant contributions to adaptation, as demonstrated by experimental evolution studies in yeast grown under highly selective conditions (Gorter *et al*., 2017; Oud *et al*., 2013). In natural yeast populations, CNVs can be highly dynamic, segregating even within very closely related populations differing only by few SNPs (Jeffares *et al*., 2017), and beneficial CNVs can quickly revert in the absence of a strongly selective environment (S.-L. Chang *et al*., 2013). This suggests that CNV serves as a genetic mechanism for rapid response and adaptation to environmental changes.

There is evidence to suggest that CNV can facilitate rapid adaptation in plants as well. CNV is prevalent within natural and domesticated plant populations (Lye & Purugganan, 2019; Zmienko *et al*., 2020) and can have a significant impact on gene expression, and consequently phenotypic variation, when it overlaps with genes or regulatory elements (Alonge *et al*., 2020). For example, a natural population of the weed species *Amaranthus palmeri* developed resistance to the herbicide glyphosate within 10 years through amplification of the EPSPS gene (Gaines *et al*., 2010). Additionally, an experimental evolution study in *A. thaliana* found CNVs occurred at significantly higher rates in lines exposed to high temperature stress (28 °C) compared to a non-stressed lines (22 °C), and that one of the duplications identified in the stressed lines overlapped with genes involved in response to temperature stress (DeBolt, 2010). However, a more recent mutation accumulation study in *A. thaliana* grown under high temperature stress failed to detect such CNVs (Belfield *et al*., 2021). Further adaptive evolution studies focusing on the role of CNVs with regards to short-term stress adaptation in plants have been lacking, in part due to the challenges associated with identifying CNVs using short read sequencing approaches (Ho *et al*., 2020). Therefore, it is unclear whether short-term adaptation through CNV is a rare occurrence or a more general phenomenon.

Here, we leverage the experimental evolution framework and recent advancements in computational variant detection methods (Wijfjes *et al*., 2019) to investigate the genetic architecture of short-term stress adaptation in plants. We employed whole genome sequencing (WGS) to detect *de novo* CNVs, SNPs, small indels, and changes in methylation in *A. thaliana* lines grown for five generations under salinity and zinc stress. Although long-read sequencing is better suited for detecting large variants such as CNVs (Sedlazeck, Rescheneder, *et al*., 2018), we opted for short-read sequencing platforms due to their lower cost, allowing for a larger number of individual plants to be analysed in each treatment. Salinity and zinc stress were specifically chosen as treatments because genes involved in conferring tolerance to these two types of stress were found to be multiplicated in plant species closely related to *A. thaliana* that are adapted to high levels of salt (Dassanayake *et al*., 2011; H.-J. Wu *et al*., 2012) and zinc (Hanikenne *et al*., 2008; Lochlainn *et al*., 2011). This study aims to elucidate the rate at which *A. thaliana* develops an adaptive response to salinity and zinc stress, the rate at which different types of genetic and epigenetic mutations accumulate during this process, and the contribution of these mutations to stress adaptation.

Results

Effect of experimental treatments on plant performance

We tested the effect of two distinct types of challenging abiotic conditions, salinity and excess zinc, on plant growth. Plants were grown for five consecutive generations under two different abiotic stress inducing treatment regimes that differed in their severity (Figure 1a). In the ‘moderate stress’ treatments the concentration of salt/zinc remained constant across generations, and we will refer to experimental populations from these treatments as either salt stress constant (SSC) or zinc stress constant (ZSC). The first generation of the ‘severe stress’ treatments started at the same salt/zinc concentration as in the moderate stress treatments, but the concentrations increased by 10% at each subsequent generation. The ‘severe stress’ experimental populations will be referred to as salt stress increasing (SSI) or zinc stress increasing (ZSI). Finally, we used separate ‘control’ treatments for salt and zinc, and these experimental populations are referred to as salt control (SC) or zinc control (ZC). These conditions were established in a pilot study where a range of different excess zinc and salinity concentrations were tested on a subset of five *A. thaliana* accessions. The accession Köln-5 (Kl-5) was chosen for this study as it is relatively susceptible to the salinity and excess zinc treatments compared to the other accessions tested, it flowered relatively early on all experimental treatments, the seed germination was close to 100% on all experimental treatments and there was resequencing data available at the start of the experiment.

Experimental populations growing under moderate and severe stress treatments had reduced growth compared to their matching control treatments, with the effect of both zinc stress treatments being more pronounced, as shown by the occurrence of chlorotic plants (Figure 1b-e) and the reduced seed yield (Figure 1f). For both salt stress treatments, the reduced seed yield could be attributed to reduced survivability of seedlings, especially during the first week after the application of the high salinity solution. The decrease in total seed yield in the zinc stress treatments instead resulted from a strongly reduced production of seeds per individual plant. Plants in the ZC treatment obtained higher total seed yields compared to the SC treatment (Figure 1f), underlining the importance of defining separate control treatments for the two stress regimes.

We measured seed ionome profiles of progeny from the fifth experimental generations and explored variation in ionome profile through principal component analysis (Figure 2a). The largest degree of variation in ionome profile (PC1 in Figure 2a) is due to differences in growth substrate: rockwool for the salinity treatments (SC, SSC, and SSI) and a peat-based soil mixture for the zinc treatments (ZC, ZSC, and ZSI). Additionally, the seed ionome profile for both zinc stress treatments mainly differs in PC2 from its respective control, while for salinity all three treatments group together (Figure 2a). Although there is overall no strong variation along PC1 and PC2 between replicates of the SC and SSC/

SSI treatments, they do show significant differences in the concentration of some individual elements, with the largest differences in sodium (Na) concentration (Figure 2b). Similarly, the seed zinc concentration is significantly higher in both zinc stress treatments (Figure 2c), where in total 14 out of the 3 elemental concentrations significantly differ between control and two stress treatments (Table S2).

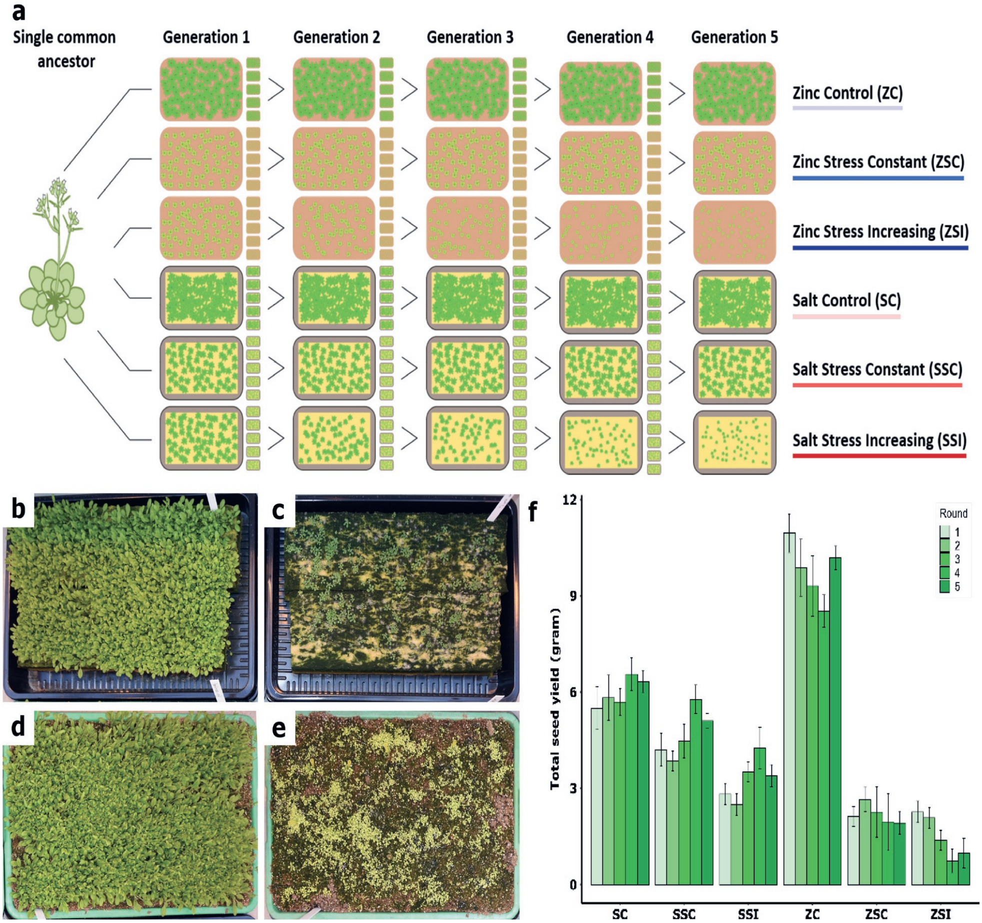


Figure 1: **Effects of experimental treatments on plant growth and total seed yield.** (a) Schematic overview of the experimental setup. Offspring from a single ancestor were used to provide the starting population for the six experimental populations that received a control, moderate or severe stress treatment (ZC, ZSC, ZSI, SC, SSC, SSI). Six independent replicate populations were maintained per treatment. Photograph of one replicate population of SC (b), SSC (c), ZC (d), ZSC (e), 12 days after sowing. (f) Total seed yield collected from six replicate populations such as depicted in (b)-(e), for each of the five generations (rounds) of the experimental treatments.

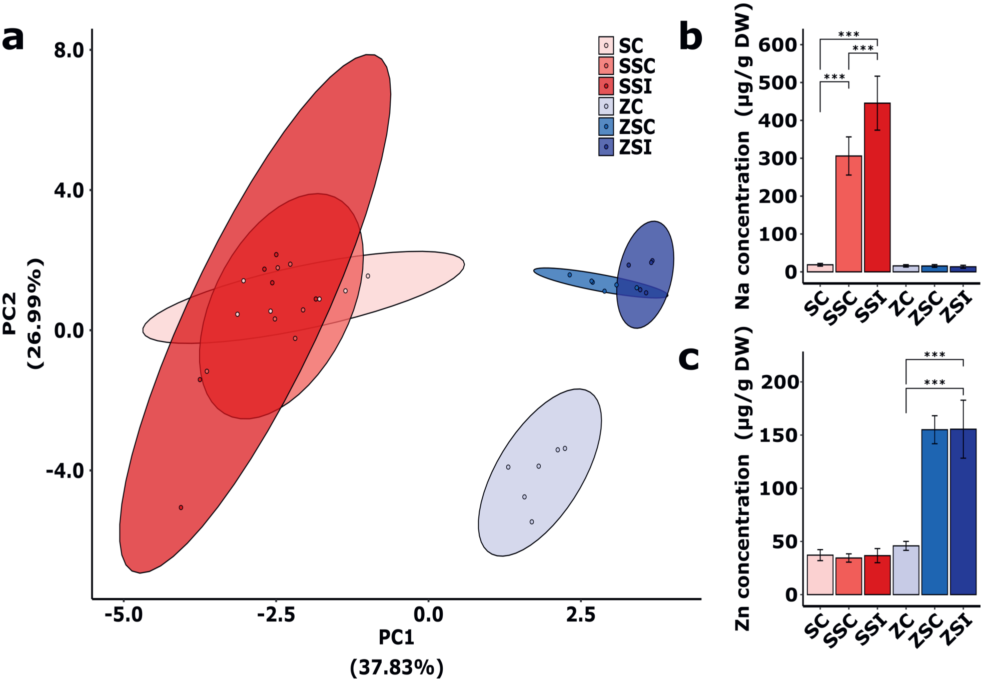


Figure 2: **Seed ionome profiles of progeny from the fifth experimental generation.** (a) Principal component analysis of ionome profiles from pools of seeds of each treatment. Seeds of each of the six replicate populations per treatment were included, with the exception of one replicate population from the severe zinc stress treatment (ZSI1). (b-c) Sodium (Na) (b) and zinc (Zn) (c) concentration of seeds. Differences were assessed using a one-way ANOVA, in which only differences within either the salt or zinc related treatments are considered. Error bars represent the standard deviation of the mean.
$^*p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

Experimental adaptation to zinc stress

To test if the experimental populations have adapted to their stress treatments, we evaluated the relative performance of offspring from the fifth generations in both the control and moderate stress treatment. We decided to focus our phenotypic evaluations exclusively on the progeny of the zinc-related treatments as the zinc stress treatments induced a larger degree of physiological stress response symptoms (e.g. plants showing strongly reduced growth and seed production) and we had only noted potential signs of adaptations in some of these replicate populations. We measured fresh weight and flowering time of progeny from the fifth generation on both the zinc control and moderate stress treatment (Figure 3a-d). Under zinc control conditions, both ZSC and ZSI populations had slightly, but significantly higher fresh weight than the ZC populations (Figure 3b) and flowered on average a few days earlier (Figure 3d). Under the moderate zinc stress conditions, there were no differences in fresh weight or flowering time between the ZC and ZSC/ZSI populations, except for one replicate population from the severe zinc stress, namely ZSI2 (Figure 3b-d). Under moderate zinc stress conditions, ZSI2 has significantly higher fresh weight than any other population (Figure 3b-c) and flowers significantly earlier (Figure 3d). The much higher fitness also results in more seeds that are produced, but as plants shattered seeds easily, seed production could not be reliably determined. Thus, overall we observe a small but noticeable effect of the zinc stress treatments on the growth of progeny in the subsequent generation under control conditions, and we find one replicate population (ZSI2) with drastically improved performance under zinc stress conditions.

Besides the increased excess zinc tolerance in replicate population ZSI2, we observed a limited number of plants with increased performance under moderate zinc stress in other ZSC and ZSI replicate populations during their fifth experimental generation. When we evaluated the growth performance of a random subset of progeny from each replicate population under the moderate zinc stress treatment, we observed one individual plant with increased zinc tolerance in replicate population ZSC3 (Figure 3e).

Limited evidence of changes in DNA methylation contributing to adaptation to zinc stress

Besides genetic variation, previous work indicated that rapid adaptation of *A. thaliana* may involve changes in cytosine methylation levels maintained for at least two to three generations in the absence of the selective environment (Schmid *et al*., 2018). We used bisulfite sequencing to investigate changes in DNA methylation between plants of stress and control treatments and compared these to changes in DNA methylation found between plants within the same control treatment. DNA was also isolated from progeny of plants from the fifth generation grown under control conditions, to exclude transient epimutations.

Differentially methylated sites were observed almost exclusively in the CpG context as opposed to the CHG and CHH context (Figures 4 and S1), consistent with previous work (Schmid *et al*., 2018). The number of CpG sites differentially methylated between plants grown under stress treatments versus those grown under control treatments did not significantly differ from the numbers observed between plants of the same control treatment (Figure 4a), nor did the genomic context in which such sites were found (Figure 4b). Nevertheless, plants grown under severe zinc stress treatment have greater variation in the number of differentially methylated CpG sites compared to plants grown under control conditions (Figure 4a), with one plant of the ZSI6 population containing the most differentially methylated sites (1554) of all sequenced plants. The plant from the ZSI2 population did not contain a significantly different number of differentially methylated sites relative to ZC plants.

We found 78 promoters (Figure 4c) and 300 transcribed regions, or ‘gene bodies’ (Figure 4d) that overlapped with differentially methylated sites exclusively in ZSI plants, suggesting that these may be involved with the treatment. Given the elusive function of gene body methylation in plants (Bewick & Schmitz, 2017), we hypothesize that epimutations that are

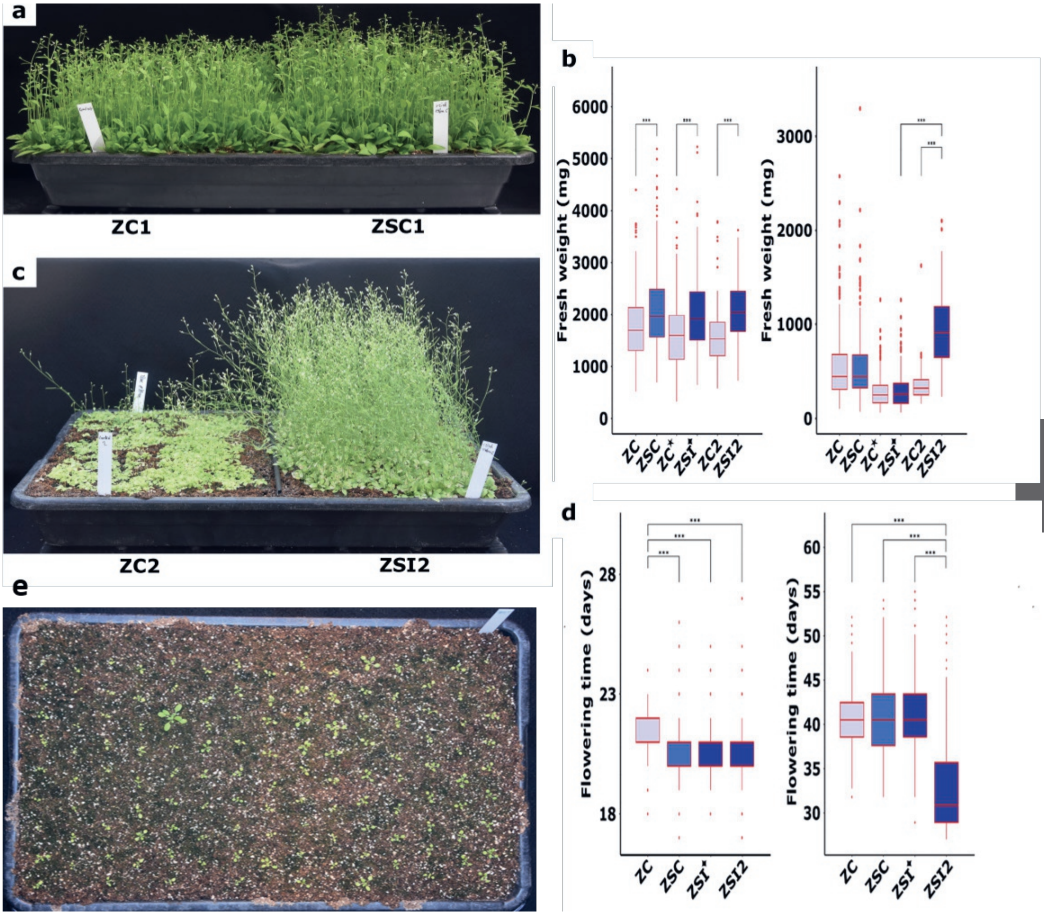


Figure 3: **Response of progeny after five experimental generations to control and moderate zinc stress.** (a) Tray containing progeny from ZC1 and ZSC1 replicate populations when grown together on the zinc control treatment. (b) Fresh weight of above ground tissue measured of 50 plants from the six replicate populations for each of the three treatments (control, moderate and severe zinc stress), grown under zinc control (left panel) or moderate zinc stress (right panel) conditions. Significance of differences was assessed using two-sample t-tests. (c) Tray containing progeny from ZC2 and ZSI2 replicate populations grown under the moderate zinc stress treatment. (d) Flowering time of offspring of different treatments grown under ZC (left panel) or ZSC (right panel) conditions. (e) Progeny from replicate population ZSC3 grown under moderate zinc stress treatment, with one offspring that is more excess zinc tolerant. Significance of differences was calculated by one-way ANOVA with a Tukey HSD posthoc test. ★ All ZC offspring, except for ZC2. ♦ All ZSI offspring, except for ZSI2. *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

the most likely to affect phenotype are those that cluster inside promoter regions of genes with a known function in stress response. Based on these assumptions, the ZSI-specific differential methylation of the promoter of the *GLYCINE-RICH PROTEIN 3* (*GRP3*) gene is the most likely to have contributed to stress adaptation, as it contains four differentially methylated sites and the associated gene has been implicated in mediating aluminium tolerance (Mangeon *et al*., 2016). However, the epigenetic modifications to this gene, or any other, do not seem to have contributed to enhanced zinc excess tolerance, as aside from the ZSI2 population no general phenotypic differentiation or adaptive response has been observed in the replicate ZSI populations.

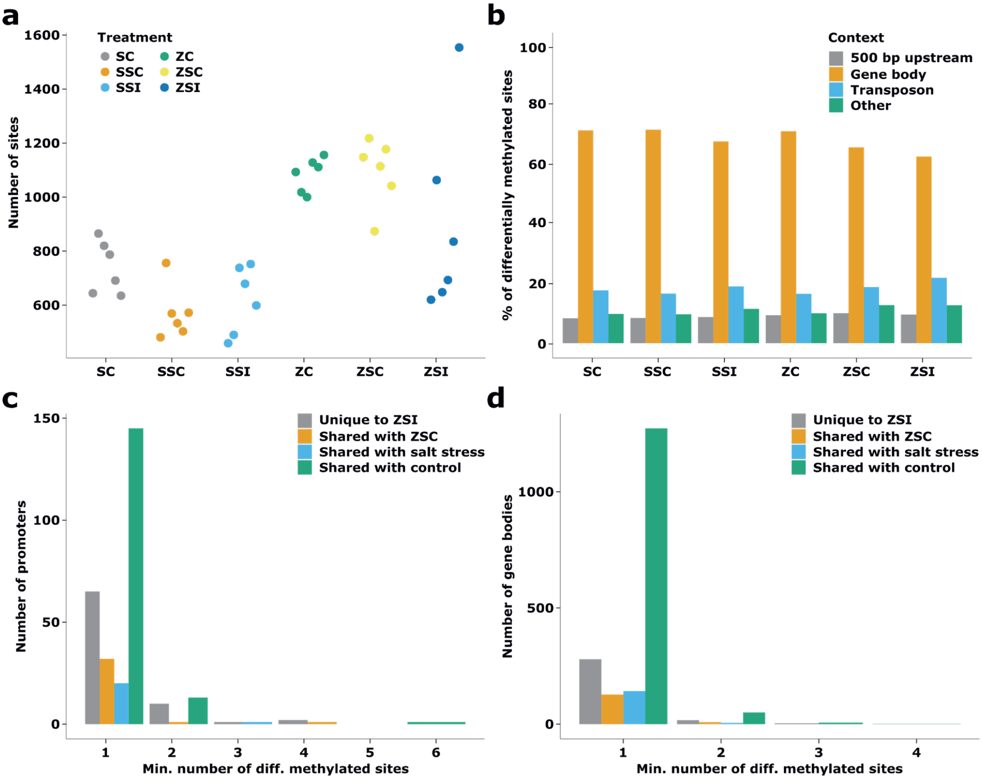


Figure 4: **Differentially methylated CpG sites in progeny of the fifth generation when grown on the zinc control treatment.** (a) The number of differentially methylated CpG sites in individual samples of each treatment. (b) Genomic context of differentially methylated sites found in the individual samples of each treatment. (c-d) The number of promoters (c), defined here as the 500 bp regions upstream of genes, and gene bodies (d) that overlap with differentially methylated CpG sites found in samples of the ZSI populations. Promoters are categorized as "Unique to ZSI" if they only overlap with differentially methylated sites of the ZSI treatment or "Shared with..." if they overlap with sites of other treatments as well. The x-axis depicts the minimum number of differentially methylated sites overlapping with the promoter resp. gene body in an individual sample.

Plants grown under increasing zinc stress contain putative adaptive genetic mutations

We set out to characterize whether the physiological and phenotypic response of plants to salinity and zinc stress affected the accumulation of spontaneous genetic mutations. Therefore, we sequenced five randomly selected progeny from each replicate population.

Plants from the ZSI treatments contained on average about a twofold higher total number of spontaneous mutations compared those from the zinc control treatment (p = 0.03, Mann-Whitney U test). Moreover, both SNPs and InDels were accumulated roughly twofold higher on average in the ZSI populations compared to the ZC populations. For plants grown under moderate zinc stress the average number of accumulated mutations is in between the ZC and ZSI populations, but that does not significantly differ from either. Plants from the SSC populations contained about 25% fewer mutations than those grown under the salt control treatment (*p* = 0.022, Mann-Whitney U test). This difference is predominantly due to a lower number of SNPs, but not of indels (Figure 5a).

Aside from the higher number of mutation observed in the ZSI populations, also a relatively larger fraction of genetic variants are shared by multiple plants from the severe zinc stress compared to the control (Figure 5b). We examined whether the mutations observed in ZSI plants likely affect gene functioning and consequently plant phenotype using variant effect prediction. A small set of SNPs and indels are predicted to have moderate to high impact (Figure 5c) and these could have a possible role in stress adaptation. The SNP predicted to have the strongest effect is a variant observed in all sequenced plants of the ZSI2 population. This variant introduces a premature stop codon in the first exon of the gene *RESPIRATORY BURST OXIDASE HOMOLOG F* (*RBOHF*), known to be involved in stress-induced production of reactive oxygen species (ROS) (Kwak *et al*., 2003; Torres *et al*., 2002). This mutation likely results in a knock-out, as it severely truncates the translated protein, resulting in the loss of all its membrane-spanning domains and calcium-binding regions that are considered vital for its function in signal transduction. Besides this variant, we found three other variants shared by multiple plants in the ZSI2 population: a 1 bp deletion in the stop codon of the gene *ZINC FINGER OF ARABIDOPSIS THALIANA 7* (*ZAT7*), a transcription factor involved in oxidative stress response and salinity stress (Xie *et al*., 2019), and two non-synonymous SNPs; one in the gene *At2g18560* (a UDP-glycosyltransferase superfamily protein) and one in *MITOCHONDRIAL SINGLE-STRANDED BINDING PROTEIN 2* (*SSB2*), which acts as a regulator of mitochondrial replication and homologous recombination (Qian *et al*., 2022).

Very few CNVs were detected overall (Table 1) with no significant difference between different treatments, similar to mutation accumulation lines grown under high temperature stress (Belfield *et al*., 2021). One of the CNVs found in a SSC plant completely overlapped with the gene *SENESCENCE-ASSOCIATED GENE 13* (*SAG13*), which is an oxidoreductase linked to senescence and previously found to be upregulated in leaves of an early-aging *A. thaliana* mutant (Schippers *et al*., 2008).

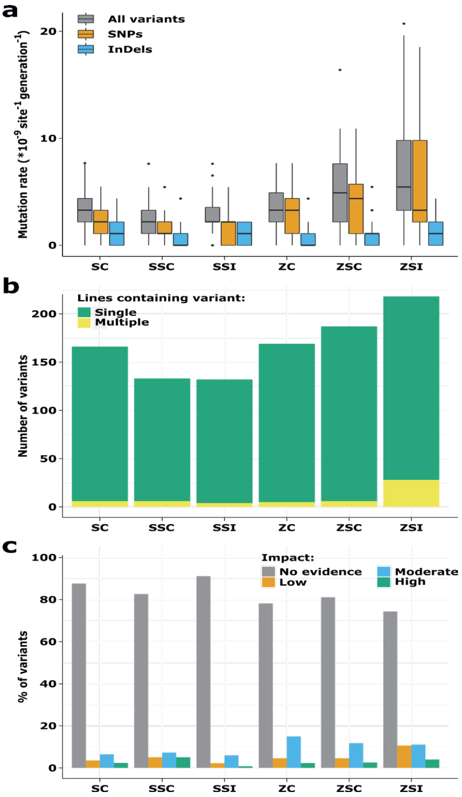


Figure 5: **Spontaneous mutations in plants from all six replicate populations per treatment after five generations of growth under the different treatments.** (a) Distribution of the mutation rate of all homozygous variants, SNPs and indels per treatment. (b) Number of samples containing variants found in each treatment. (c) Impact of SNPs and indels in each treatment, as computed by variant effect prediction (VEP).

Table 1: **Number of CNVs detected in generation 5 of plants grown under each population.**

Population	Number of CNVs	Population	Number of CNVs
SC	4	ZC	3
SSC	2	ZSC	2
SSI	2	ZSI	0

As the founder accession used in all experiments (Kl-5) is different from the one used to generate the *A. thaliana* reference genome (Col-0), we verified that using Col-0 as a reference genome to detect *de novo* variants did not result in missing putative adaptive variants. 98% of the reads of two offspring samples of the founder accession (generation 1) mapped to the Col-0 reference genome (Table S3). Of the unmapped reads, 41 to 51% could be aligned to chromosome-level assemblies of seven different *A. thaliana* accessions, covering a small number of genes of each (Table S3). These genes are associated with various Gene Ontology terms, although none directly involved with adaptation to high salinity or zinc adaptation (Table S4). They could be grouped into 136 homology groups, each representing a gene of Kl-5 missing from the Col-0 reference genome. Given this small number of homology groups, the limited number of generations propagated in all experiments, and the observed low spontaneous *de novo* mutation rate, it is unlikely that we missed putative adaptive variants by mapping to the Col-0 reference genome only.

The genetic basis of zinc tolerance in the ZSI2 population maps to a premature stop codon in *RBOHF*

To further establish the (epi)genetic cause of zinc-tolerance in the ZSI2 replicate population, we tested if all plants from the ZSI2 population were zinc-tolerant by growing a random sample on the moderate zinc stress treatment in a grid format to evaluate individual performances. This demonstrated the presence of two phenotypically distinct classes, with approximately 15% of plants appearing equally excess zinc-sensitive as other ZC, ZSC and ZSI populations, and about 85% of progeny that are significantly more tolerant to the moderate zinc stress treatment (Figure 6a). Next, we evaluated the performance of progeny from individual zinc-tolerant ZSI2 lines under moderate zinc stress treatment and observed that all their progeny are equally zinc-tolerant as their parent. Moreover, when zinc-tolerant ZSI2 lines were propagated for two consecutive generations under zinc control conditions, and then tested to moderate zinc stress, their progeny remained zinc-tolerant. Together this demonstrates a genetic, rather than epigenetic or physiological, cause of zinc-tolerance.

To determine the genetic cause of zinc-tolerance, we crossed a zinc tolerant ZSI2 line (as a father) with Col-0, which is sensitive to the moderate zinc stress treatment, to produce an F_1. This F_1 exhibited similar sensitivity to the moderate zinc stress treatment as Col-0, indicating the trait is genetically recessive. The F_2 progeny segregated for zinc-tolerance in a ratio most close to 5:1 sensitive to tolerant (1168 sensitive vs 238 tolerant plants). This ratio best resembles may point to a single locus as it is quite close to a 3:1 segregation ratio, or may point to a segregation of two genetically linked genes that are jointly required for zinc-tolerance. Bulked segregant analysis was performed, comparing pools of DNA obtained from zinc-sensitive and zinc-tolerant plants, in order to map the genetic locus responsible for zinc tolerance. One quantitative trait locus (QTL) is mapped at Chr. 1 (Figure 6b). The top of the peak of this QTL mapped just over 1100 bp away from the *de novo* mutation which causes a premature stop codon in the *RBOHF* gene, and which is shared by all five sequenced ZSI2 plants. Additionally, the QTL spanned two other *de novo* mutations that were shared among respectively three and five of the sequenced ZSI2 individuals. The first of these mutations is located at approximately 16 Mb in the intergenic space between two pseudogenes, while the second is located near 26 Mb and 395 bp upstream of the start codon of *At1g69060* (Chaperone DnaJ-domain superfamily protein) (Figure 6b). The *de novo* mutation in *RBOHF* appears as the most likely mutation affecting the zinc tolerance, as this loss-of-function mutation has the highest predicted impact on gene function and completely co-segregates with the phenotype.

To validate if a loss-of-function mutation of the *RBOHF* gene could increase zinc tolerance we tested the Col-0 *RBOHF*-F3 knockout line (Torres *et al.*, 2002) for its response to moderate zinc stress treatment.

Col-0 *RBOHF*-F3 does not exhibit increased zinc tolerance and is equally sensitive to zinc stress as Col-0 and ZC plants (Figure 6c).

Whole rosette ionome analysis was conducted on plants grown on the control and moderate zinc stress treatment. The average rosette Zn concentration ranged between 20 to 36 µg/g DW under the zinc control treatment (Figure 6d), and between 1133 and 1557 µg/g DW under the moderate zinc treatment (Figure 6e). The Zn concentration in ZSI2 was significantly lower compared to ZC1 (on average 1133 versus 1486 µg/g DW respectively, p = 0.037) on the moderate zinc stress treatment, but did not differ significantly on the zinc control treatment. In contrast, no significant differences are observed between Col-0 and Col-0 *RBOHF*. Together, this indicates that a loss-of-function mutation in *RBOHF* is either not causal or this mutation is not sufficient on its own to cause zinc tolerance.

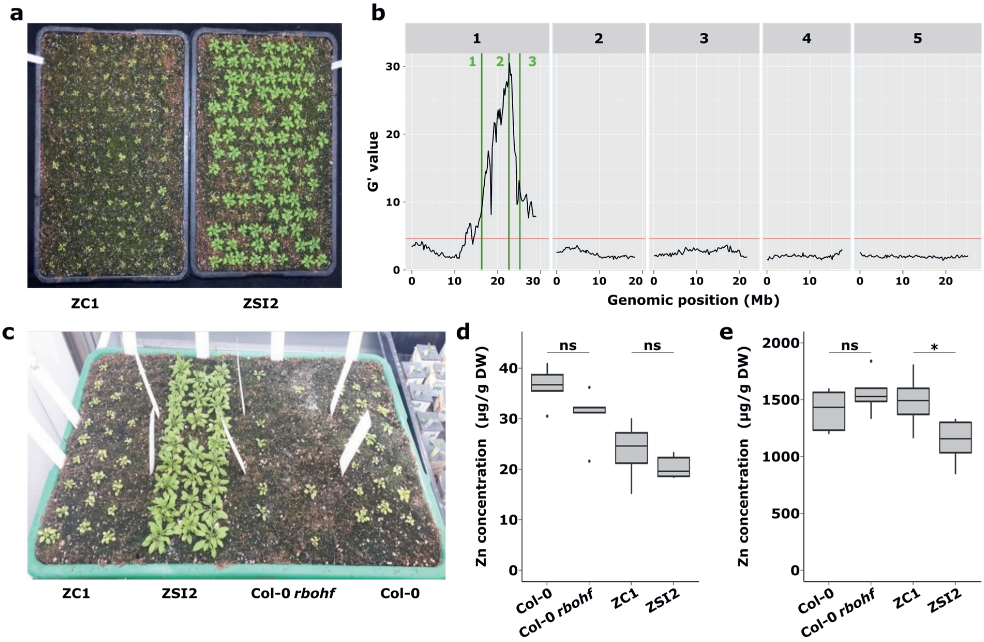


Figure 6: **Mapping of the causal mutation(s) underlying excess zinc-tolerance.** (a) Plants from the ZC1 and ZSI2 replicate populations grown on moderate zinc stress. The tray containing ZC1 individuals serves as a control and comparison to evaluate the phenotypic effects of the moderate zinc stress treatment. (b) Bulked segregant analysis reveals one QTL for the association with zinc tolerance among the F_2 progeny of the cross between Col-0 and ZSI2. Green lines with a number (1-3) refer to three de novo mutations that were detected in the ZSI2 population and were shared by at least three out of five sequenced individuals. Mutation (1) is located between two pseudogenes, mutation (2) confers a premature stop codon in *RBOHF* and mutation (3) is located just upstream of *At1g69060*. (c) Validation experiment to test the Col-0 *rbohf* mutant, grown on moderate zinc stress, with ZC1, ZSI2 and Col-0 as positive controls. Rosette Zn concentration for the same four genotypes when grown under (d) control treatment and (e) moderate zinc stress treatment.

Discussion

Certain types of environmental conditions imposing stress on plants have been shown to affect the frequency and spectrum of spontaneous mutations in plants (Belfield *et al*., 2021; C. Jiang *et al*., 2014; Z. Lu *et al*., 2021). However, the speed at which plant populations can adapt to stress and the genetic mechanisms underlying such adaptation are poorly understood. Little is known about the contribution of copy number variants, a class of genetic variants previously implicated with rapid stress adaptation of plants (DeBolt, 2010; Gaines *et al*., 2010), in particular. Therefore, we assessed the extent to which *A. thaliana* can adapt to salinity and zinc stress within a few generations and the genetic mutations underlying such adaptation. Plants showed an obviously visible physiological and phenotypic response after being exposed to salinity or excess zinc stress for five generations and seeds of such plants had respectively a higher sodium or zinc concentration. One particular replicate population grown on the severe zinc stress treatment, ZSI2, showed strong adaptation, while in independent replicate populations we observed a limited number of individual plants with increased zinc tolerance. Although these have not yet been followed up, it demonstrates that populations of *A. thaliana* can rapidly become tolerant to this treatment.

Contrary to previous mutation accumulation (MA) experiments using *A. thaliana* (Belfield *et al*., 2021; C. Jiang *et al*., 2014; Z. Lu *et al*., 2021), we do not find an overall trend of stress increasing the rate of spontaneous mutations. Compared to the control treatment, plants under severe zinc stress treatment acquired, on average, slightly more than a twofold increase in the numbers of mutations. In contrast, the intermediately stressed plants accumulated, on average, approximately 40% more mutations, but this difference is not statistically significant from plants growing under control conditions. For plants from the zinc-adapted ZSI2 replicate population, the number of *de novo* mutations ranged from roughly 3.5 to 5.5 times higher than the average of the zinc control, which is primarily due to the strong selection on fitness, in this case of a single genotype containing a relatively high number of mutations. In contrast, no increase in mutation rate is observed in response to the salinity treatments. In fact, the total number of mutations was slightly lower in the intermediate salinity treatment compared to its control. In a similar study, Jiang *et al*. (2014) applied salinity stress (125 µM NaCl) for 10 consecutive generations to generate mutation accumulation (MA) lines (3 in control and 3 in high salinity) and reported a twofold higher mutation rate and increased levels of epimutations in the salt-treated lines. These observations by Jiang *et al*. (2014) align with findings of increased mutation rates in *A. thaliana* suspension cultured plant cells when cultured under high salinity conditions (X. Zhu *et al*., 2021).

The discrepancies in response to salinity stress between our study and those of Jiang *et al*. (2014), and to a lesser degree Zhu *et al*. (2021), are challenging to explain. The MA approach used by Jiang *et al*. (2014) should be more effective at observing all spontaneous mutations, whereas in our experimental setup, selection against deleterious mutations may have reduced the total number of observed mutations. Additionally, our experimental design introduces some complexities as the level of selection is not solely dependent on higher salinity or zinc levels, but is also influenced by several other experimental factors, some of which are treatment-specific. For instance, due to the reduced seed production of plants under the highest levels of stress, the population size was adjusted in the severe-stress treatments to ensure enough seeds to start a subsequent generation. As every treatment was initiated with the same population size, the faster growth and larger size of plants in the control treatment resulted in greater competition for light and other resources compared to any stress treatment. It is uncertain how these differences in selective pressure have affected the results in the salinity treatments, as such effects would also be expected to influence the zinc treatments.

Overall, when comparing the effects of treatment on fitness in Figure 1f with the mutation rates in Figure 5a, the results suggest that mutation rates may only increase once a certain threshold level of stress is reached. It is possible that the salinity treatments we applied were not sufficiently stressful to reach that level. A study on *A. thaliana* MA lines exposed to three temperature treatments (control, moderate warming and high temperature) for 22 consecutive generations yielded similar results as our zinc treatments, with significantly higher mutation rates at the highest level of temperature stress and intermediate mutation rates at the moderate warming (Z. Lu *et al*., 2021). Two separate *A. thaliana* MA studies also confirmed higher mutation rates in response to high temperatures (up to 29 ^{0}C and 33 ^{0}C) (Belfield *et al*., 2021; Yadav *et al*., 2022), but not to low temperatures (16 ^{0}C) (Belfield *et al*., 2021). Although the impact of multigenerational stresses on mutation rates in plants has not been extensively studied, a general trend is observed where stress increases the mutation rate once it reaches a certain minimal threshold level.

The validation of the causal gene(s) that confer zinc tolerance is still a work in progress. The loss-of-function mutation in *RBOHF* is likely to be involved, in zinc tolerance, but does not completely explain the tolerant phenotype. Only very few spontaneous mutations that occurred on chromosome 1 are shared between tolerant plants. Among these mutations, only the one in *RBOHF* shows a perfect correlation with zinc tolerance, as 100% of the sequencing reads from the tolerant pool in the BSA contained this mutation. Furthermore, the segregation ratio among F_2 progeny either suggests the involvement of one single recessive locus (in the case of a 3:1 segregation ratio), which aligns well with the loss-of-function *RBOHF* allele. Alternatively, the ratio of 5:1 suggests the involvement of two linked genes with an estimated genetic distance of approximately 10 cM. In this case it is highly likely that *RBOHF* is one of those genes involved. The presumed knockout mutant of *RBOHF* in Col-0 did not confirm the hypothesis that a loss-of-function allele of *RBOHF* by itself can directly improve zinc tolerance. Currently, several possible explanations for the lack of validation are being considered. Firstly, the Col-0 *RBOHF* mutant we used may not be a complete knock-out mutant. The mutant used in the validation experiment is a dSpm transposon insertion line with an insertion in the first exon (Torres *et al*., 2002). Although RNA gel blot analysis and RT-PCR suggest that this line is a knockout, the authors also reported the presence of several different chimeric transcripts transposon parts and part of the gene and noted that these '*presumable would give rise to nonfunctional proteins*' (Torres *et al*., 2002). Therefore, the tested Col-0 *RBOHF* may not represent a full knockout mutant, contrary to the mutant *RBOHF* allele in Kl-5.

Secondly, there may be a second gene genetically linked to *RBOHF* that is involved, and allelic variation between Col-0 and Kl-5 exists for this gene. If Kl-5 has a different allele that is a prerequisite for zinc tolerance, loss-of-function of *RBOHF* alone may not confer zinc tolerance in the Col-0 background. To address these possibilities, CRISPR-Cas9 knockouts of *RBOHF* are currently being generated in both the Col-0 and Kl-5 backgrounds. This mutant analysis should help determine whether loss-of-function of *RBOHF* alone is sufficient for increased zinc tolerance in Kl-5 or if a second (or even more) gene(s) is/are involved. If a second gene is involved, it should be located on chromosome 1 nearby *RBOHF* because of the segregation ratio observed among the F_2 progeny derived from the cross between ZSI2 and Col-0, and because the BSA only detected a single QTL. To further examine this potential second gene, a BSA will be performed on the F_2 progeny resulting from a cross made between Col-0 *RBOHF* and ZSI2. In this case, only the second locus should segregate among the F_2 progeny.

Thirdly, the most challenging possibility to resolve would be if the peak in the BSA has indicated *RBOHF* by chance, but in reality, an unknown spontaneous mutation located nearby is the causal factor and has evaded detection by the variant calling approach. Therefore, if the other approaches suggest that *RBOHF* is not involved, a long-read *de novo* assembly of a zinc tolerant line should be constructed and compared to the ancestor to identify previously undetected genetic variations.

The mutation in *RBOHF* may be beneficial under excess zinc stress conditions, as it has been proposed that different ROS signatures help to tailor stress acclimation responses of plants to the specific stress they encounter (Choudhury *et al*., 2017). ROS can be generated in the apoplast by cell wall peroxidases and plasma membrane-localized flavin-containing NADPH oxidases (NOX), which belong to the *RESPIRATORY BURST OXIDASE HOMOLOG* (*RBOH*) family in plants (H. Huang *et al*., 2019; Kadota *et al*., 2015). Plants possess multiple members of this family, each serving mostly dedicated functions (Torres & Dangl, 2005). *RBOHF* is involved in fine-tuning ROS levels in response to various (a)biotic stresses and plays a pivotal role in plant plastic responses to their environment (Kaur *et al*., 2014). The effects of ROS signalling can extend over long distances through the propagation of ROS waves, leading to systemic responses throughout the entire plant (Fichman & Mittler, 2020). These ROS waves are self-propagating and rely on the activity of RBOHD and RBOHF proteins (Fichman & Mittler, 2020; Zandalinas *et al*., 2020). In response to specific types of stress, apoplastic calcium ions (Ca^{2+}) activate RBOHD and RBOHF proteins, resulting in the production of hydrogen peroxide in the apoplast (Mhamdi & Van Breusegem, 2018; Mohanta *et al*., 2018; Zandalinas *et al*., 2020). Hydrogen peroxide, being relatively stable, can accumulate in the apoplast and then be transported from cell to cell via aquaporins located in the cell membrane (Bienert *et al*., 2007; Mhamdi &

Van Breusegem, 2018). As neighbouring cells detect the influx of hydrogen peroxide, they respond in a similar manner, allowing the signal to propagate as an auto-propagating wave throughout the entire plant (Mhamdi & Van Breusegem, 2018). ROS waves can be induced by various types of abiotic and biotic stress, and they are considered a general yet essential signal that alerts cells and tissues to impending stress, rather than a signal that triggers specific responses based on the specific type of stress (Fichman & Mittler, 2020). *RBOHF* is thus a central hub in a complex regulatory network of stress signalling pathways (J. Han *et al*., 2019). However, the specific role of *RBOHF* to zinc tolerance has to be assessed further.

If *RBOHF* is indeed involved in the zinc tolerant phenotype of ZSI2, this may indicate that zinc tolerance results from a lack of ROS signalling along with a slight reduction in zinc accumulation. In one of the experiments, accidentally fifty times higher concentrations of certain components (KI, Na_2MoO_4*$2H_2O$, $CoCl_2$ and $CuSO_4$*$5H_2O$) were added to MS medium when preparing vertical agar plates. Zinc-tolerant ZSI2 plants exhibited increased levels of tolerance to these components as well (Figure 7). These results suggest that the tolerance observed is not specific to zinc but likely represents a more general stress response that is affected, and that can have beneficial and deleterious effects depending on the specific environment. *A. thaliana* may sense the high soil zinc concentration as an imminent stress and react with a systemic stress response that signals the plant to stop growth. Such a systemic stress response may follow from ROS waves, or may involve jasmonic acid (JA) signalling. Many (a)biotic stresses, including several heavy metals increase JA levels (Fonseca *et al*., 2009; Maksymiec *et al*., 2005). Increased JA levels inhibit plant growth by suppressing cell division and proliferation (Noir *et al*., 2013; Y. I. Zhang & Turner, 2008), but the activity of JA signalling depends on the activity of *RBOHD* and *RBOHF* (Maruta *et al*., 2011). By removing a key regulator in the form of *RBOHF*, this response may be abolished. That would then also imply that the 'real' phytotoxic effects of zinc would be lower than is currently known, as zinc tolerant ZSI2 plants still accumulate zinc concentrations, averaging around 1100 µg g^{-1} dry weight. Although the zinc-tolerant ZSI2 genotype accumulated approximately 25% lower zinc concentration compared to the ZC1 population, it seems unlikely that this reduction in zinc concentration can explain the significantly improved performance and fitness. The observed zinc concentration in ZSI2 is well above the phytotoxic threshold of 300 µg Zn g^{-1} dry weight, which is considered harmful to most plants (Broadley *et al*., 2007). Transcriptomic analysis may further elucidate what processes are changed in the ZSI2 plants and may aid in detecting the causal gene(s) as well.

We conclude that *A. thaliana* can rapidly adapt to environmental stress through *de novo* mutations within a short amount of time, given a highly selective environment. Importantly, it strengthens this phenomenon beyond earlier examples involving herbicide resistance (Gaines *et al*., 2010) and tolerance to high temperature (DeBolt, 2010). In contrast to the experiment by DeBolt, 2010), we did not find evidence that rapid generation of *de novo* CNV has played an important contribution to rapid adaptation. We expect our findings to be of immediate interest for conservation biology and crop breeding, as they imply that *de novo* mutations can quickly introduce the phenotypic divergence necessary to facilitate adaptation of plants to harsh environments and the generation of elite crop cultivars through breeding.

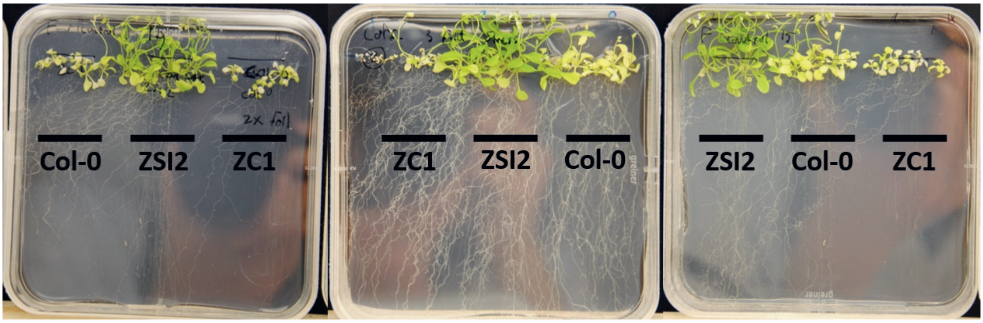


Figure 7: **Vertical plate essay using Col-0, zinc tolerant ZSI2 and progeny from a population adapted to the zinc control treatment (ZC1).** Plate on the left is a control, with ½ strength MS agar. The plate in the middle and right contain ½ strength MS agar with 50x times higher concentrations of KI, $Na_2MoO_4 \cdot 2H_2O$, $CoCl_2$ and $CuSO_4 \cdot 5H_2O$.

Materials and Methods

Growth of control and stressed lines

For each treatment, the experiment was initiated using seeds from a well-fertilised single plant of the Kl-5 accession (germplasm ID: CS76528). To start the treatment, 50 mg of seeds (approximately 2500 seeds (Jako *et al*., 2001)) per replicate population were sown and then stratified at 4 °C in a climate-controlled chamber for 3 days. Following stratification, the plants were grown in a climate controlled greenhouse with ambient CO_2 levels, with temperatures set at 20/18 °C (day/night), 70% relative humidity, and providing natural light supplemented to a minimum of 16h of light at 125 $\mu mol \cdot m^{-2} \cdot s^{-1}$. Six replicate populations were used for each experimental treatment. The seeds were collectively harvested and stored for each experimental replicate once all plants within that replicate had completed their life cycle. After harvesting the seeds, the total seed yield per experimental replicate was mixed, and 50 mg of the seed mixture was used to initiate the next generation.

For the salinity control and stress treatments, rockwool slabs (Grodan Rockwool Group, L 100 cm × W 15 cm × H 7 cm) were used and cut in half. Both halves were placed directly adjacent to each other in a plastic tray, providing a growth surface of 50 cm × 30 cm. Slabs were pre-soaked in a nutrient solution designed for *A. thaliana* (1.7 mM NH_4^+, 4.5 mM K^+, 0.4 mM Na^+, 2.3 mM Ca^{2+}, 1.5 mM Mg^{2+}, 4.4 mM NO_3^-, 0.2 mM Cl^-, 3.5 mM SO_4^{2-}, 0.6 mM HCO_3^-, 1.12 mM PO_4^{3-}, 0.23 mM SiO_3^{2-}, 21 µM Fe^{2+} (chelated with 3% diethylene triaminopentaacetic acid), 3.4 µM Mn^{2+}, 4.7 µM Zn^{2+}, 14 µM BO_3^{3-} and 6.9 µM Cu^{2+} at pH 7, EC 1.4 mS cm^{-1}) which will be referred to as Hyponex solution. Three different treatments were applied:

1: Salt control treatment: plants were watered with the regular Hyponex solution throughout the experiment. Replicate populations originating from this treatment are referred to as Salt Control (SC).
2: Moderate salt stress treatment: plants were allowed to germinate on the regular Hyponex solution, after which, one week after sowing, 1 L of Hyponex with an increased concentration of NaCl was added to the tray, raising the NaCl concentration to 175 mM. Replicate populations originating from this treatment are referred to as Salt Stress Constant (SSC).
3: Severe salt stress treatment: This treatment was identical to the moderate salt stress treatment in the first generation, but after the first generation, the NaCl concentration was increased by 10% in each consecutive generation. Replicate populations originating from this treatment are referred to as Salt Stress Increasing (SSI).

A fertilised peat-based soil mixture was used as a substrate for the zinc control and zinc stress treatments. Seeds were sown directly on treated soil. We used plastic growing tray with a surface area of 37 × 56 cm. Plants were watered with tap water. Treatments were prepared using 15 L of the potting mixture to which 5 L of treatment solution was added and thoroughly mixed. Three different treatments were applied:
1: Zinc control treatment: Plants were grown on the peat-based soil mixture, to which 5 L of demi water was added. Replicate populations originating from this treatment are referred to as Zinc Control (ZC).
2: Moderate zinc stress treatment: Plants were grown on the peat-based soil mixture, to which 10 g of $ZnSO_4 \cdot 7\ H_2O$ was dissolved into 5 L of demi water, and added to the peat-based soil mixture. Replicate populations originating from this treatment are referred to as Zinc Stress Constant (ZSC).
3: Severe zinc stress treatment: This treatment was identical to the moderate zinc stress treatment in the first generation, but after the first generation, the amount of Zn dissolved into the 5 L demi water was increased by 10% in each consecutive generation. Replicate populations originating from this treatment are referred to as Zinc Stress Increasing (ZSI).

Determination of phenotypes

Total seed yield. After each generation in each treatment, the total seed yield from all plants in a tray was weighed. The seeds were harvested from all the plants in the tray at a single time point when they had fully ripened. Before weighing, the seeds were thoroughly cleaned to remove debris.

Fresh weight. Fresh weight was measured of offspring from the fifth generation of each zinc treatment. The same experimental setup was used as described for the zinc control and moderate zinc stress treatments, with two adjustments. Firstly, the trays with treatment soil were divided into two equal halves, each measuring 37 × 28 cm. One half contained the progeny from ZC populations, while the other half contained the progeny from either ZSC or ZSI populations. The second adjustment involved using half of the number of seeds (25 mg) compared to the original experimental setup to achieve the same growing density as in the five treatment generations. Fresh weight was measured for 50 plants from the control and stress treatments. This measurement was done using a systematic sampling approach, where ten uniformly distributed sampling positions were marked, with five plants per position. The fresh weights were measured 20 days after sowing, at a stage where all of the measured plants were bolting.

Flowering time. Flowering time was measured for progeny from the fifth generation of three replicate populations of each zinc treatment: ZC1, ZC2, ZC3, ZSC1, ZSC2, ZSC3 and ZSI1, ZSI2, ZSI3. Flowering time measurements were taken as the day when the first flower of each individual plant was completely opened. Flowering time measurements were taken for plants grown on both zinc control and moderate zinc stress treatment soils. For the zinc control treatment, flowering time was measured by growing individual plants (N = 104 per replicate population) in separate pots, using the same peat-based potting mixture as used in the zinc control treatment. For the moderate zinc stress treatment, flowering time was measured in the same conditions as the moderate zinc stress treatment. However, plants were sown in a grid format, with 128 plants per tray. Two replicate trays were used to phenotype the progeny from each of the replicate populations.

Ionome analysis. Total elemental concentrations (ionome profile) was measured by inductively coupled plasma mass spectrometry (ICP-MS) from either seeds, and in a separate experiment from 3-week-old rosettes, as described by Pauli *et al*. (2018). Principal component analysis of seed ionome profiles was performed using the prcomp function in R. Results were visualized using ggplot2 (Wickham, 2016) (v3.3.2).

DNA extraction, library preparation, and sequencing

Genomic DNA was extracted from progeny of the fifth generation of the experimental treatment, and from two offspring plants of the Kl5 plant used to initiate all treatments. Plants were grown hydroponically on rockwool blocks (Grodan Rockwool Group, 40 × 40 mm in size) pre-soaked in Hyponex. Genomic DNA was extracted from inflorescences (open flowers and above) of single plants. Material was frozen in liquid nitrogen, ground to a fine dust, and incubated in 300 µL 2x CTAB buffer (2 % CTAB, 1.4 M NaCl, 100 mM Tris, 20 mM EDTA, pH 8) for 30 minutes at 65 °C. An equal volume of chloroform was added, mixed, centrifuged (3250 rpm for 15 min), and supernatant was collected. DNA was precipitated by adding an equal volume of ice-cold isopropanol, incubated overnight at -20 °C, followed by centrifugation (3250 rpm for 15 min). The precipitate was washed twice with 70 % ethanol and air dried. DNA was dissolved in milliQ water and treated with RNase (Promega) for 30 minutes at 37 °C.

DNA library preparation and sequencing were performed at Novogene (Cambridge, United Kingdom). After discarding samples that failed to pass quality control checks, regular DNA sequencing was performed of at least 30 samples per treatment and bisulfite sequencing was performed of six samples per treatment (Table S1). Illumina short-read libraries for regular DNA sequencing were prepared using the NEBNext Ultra II DNA Library Prep Kit for Illumina. The same kit was used to prepare libraries for bisulfite sequencing, after treating DNA samples with bisulfite (EZ DNA Methylation Gold Kit of Zymo Research). All libraries were sequenced using the Illumina Novaseq 6000 platform, yielding at least 5.2 Gb of paired-end (2x150 bp) data per sample.

Trimming and aligning reads of regular DNA sequencing libraries

Reads of the regular DNA sequencing libraries were trimmed using Trim Galore (version 0.6.5), a wrapper of Cutadapt (Martin, 2011) (version 1.18), with default parameters. This step clipped sequences that matched at least 90% of the total length of the standard Illumina adapter AGATCGGAAGAGC. In addition, it trimmed bases from the 5′ and 3′ ends of reads if they had a Phred score of 20 or lower. Reads shorter than 20 bp after trimming were discarded.

Trimmed reads were aligned to a modified version of the *A. thaliana* Col-0 reference genome (TAIR10, European Nucleotide Accession number: GCA_000001735.2), which contains an improved assembly of the mitochondrial sequence (Sequence Read Archive accession number: BK010421) (Sloan *et al*., 2018), using bwa mem (H. Li, 2013) (version 0.7.17) with default parameters. The resulting alignment files were sorted and indexed using samtools (H. Li *et al*., 2009) (version 1.9). Alignment files of libraries generated from the same line were merged using Picard MarkDuplicates (https://broadinstitute.github.io/picard/) (version 2.22.2). Picard MarkDuplicates was also used to mark duplicate read pairs, using an optical duplicate pixel distance of 2500, which is appropriate when working with the patterned Illumina flowcells of the Novaseq 6000 platform.

Single nucleotide polymorphism (SNPs) and small indel calling

SNPs and small indels were called in each sample using GATK (Van der Auwera *et al*., 2013) (version 4.0.2.1) HaplotypeCaller, allowing a maximum of three alternate alleles at each site. Samples were jointly genotyped using the GATK modules GenomicsDBImport, CombineGVCFs, and GenotypeGVCFs with default parameters. This step generated three different VCF files: one containing the calls of the nuclear genome, one containing calls of the mitochondrial genome, and one containing calls of the chloroplast genome. We implemented a custom filtering pipeline to filter these sets to a high-confidence set of SNP and indel calls (see Supplementary Methods). Calls were considered to correspond to *de novo* mutations if they were found within plants of a single treatment only and absent from the two plants of generation 1. Otherwise, it was assumed that the mutations were already in the Kl-5 accession from which all lines were derived.

Copy number variation calling

We called CNVs, defined here as deletions, duplications, and insertions of at least 50 bp, by applying the Hecaton workflow (Wijfjes *et al*., 2019) (version 0.5.0) with default parameters to the alignment files of each sample. CNVs called in each individual sample were genotyped and combined into a single VCF file (see Supplementary Methods for details). CNV calls were considered *de novo* mutations if they were only found in lines of a single treatment, had a read depth lower than 0.5 (deletions) or higher than 1.5 (tandem and dispersed duplications) in such lines, had a read depth lower than 0.7 (deletions) or higher than 1.3 (tandem and dispersed duplications) in maximally three lines of other treatments, and were not called in the two plants of generation 1. While these filters are stringent, we believe them necessary to reduce the number of false positive *de novo* CNV calls, as regions are known to show changes in read depth even in the absence of CNV (Pedersen & Quinlan, 2019). Insertions were not further taken into account in downstream analyses, as they are difficult to reliably detect using our workflow (Wijfjes *et al*., 2019) and with short reads in general (Ho *et al*., 2020).

Calculation and comparison of mutation rates

Mutation rates of variants were computed in each sample by dividing the total number of homozygous variants by the number of bases having a coverage between 5x and 150x, assuming that variants could be reliably called at such positions. This mutation rate was then divided by the number of generations to compute the mutation rate per bp per generation. The coverage of each base in each sample was computed using mosdepth (Pedersen & Quinlan, 2018) (version 0.2.9). Differences between mutational rates were tested for statistical significance by applying the Mann-Whitney U test to the distributions of the number of variants found in samples of two different treatments. We considered rates to be significantly different if tests yielded a *p*-value below 0.05.

Inspecting non-reference genomic content of the founder accession

Reads of two offspring plants of the Kl-5 founder accession were mapped to the Col-0 reference genome and seven chromosome-level assemblies of different *A. thaliana* accessions representing the plant's global distribution (W.-B. Jiao & Schneeberger, 2020) using PanTools (Sheikhizadeh *et al*., 2016) (branch "pantools_v3" from https://git.wur.nl/bioinformatics/pantools; command: "pantools map -am 2"). Reads that did not map against the Col-0 assembly were extracted and realigned to the other seven assemblies using the same parameters. Genes of non-reference assemblies were considered "covered" if they had an average read coverage of at least 10 and at least 90% of their length had a coverage of at least 1. Protein sequences of covered genes were assigned Gene Ontology terms using InterProScan 5.50-84.0 (Jones *et al*., 2014). Genes of all 8 genomes were grouped into homology groups with the command "pantools optimal_grouping", using a similarity score of 75%.

Differential methylation calling

Bisulfite reads were trimmed using Trim Galore with default parameters. The methylseq workflow (https://github.com/nf-core/methylseq) was used to align the trimmed reads to the same modified *A. thaliana* Col-0 reference genome as the one used for the regular DNA sequencing reads and to call methylated sites with Bismark (Krueger & Andrews, 2011) (version 0.22.3). The generated coverage files were converted to cytosine reports for all three sequence contexts (CpG, CHG, CHH) using the coverage2cytosine script of Bismark. These reports were provided as input to the R package methylKit (Akalin *et al*., 2012) (version 1.16.0) to test whether sites were significantly differentially methylated between individual lines using Fisher's exact test with default parameters of methylKit. Sites were only tested if they had a coverage of at least 10x in all sequenced lines.

To identify changes in DNA methylation potentially involved in stress adaptation, we tested, for each site in each sample grown under a stress treatment, whether it was differentially methylated compared to all of the samples of the matching control treatment, performing separate tests for each pair (e.g. stress-1 vs. control-1, stress-1 vs. control-2 etc.). We used the same procedure to test whether sites were significantly differentially methylated in a control sample compared to all other samples of the same control treatment, to get an idea of the number of differential methylated sites within a single sample that can be expected under standard conditions (e.g. control-1 vs. control-2, control-1 vs. control-3 etc.). In both cases, *p*-values of separate pairwise tests were combined using Stouffer's method, resulting in a single one-vs-all *p*-value for each site in each sample (e.g. stress-1 vs. all control samples). Combined *p*-values were separately corrected for multiple testing in each of the three sequencing contexts (CpG, CHG, CHH) using the Benjamini-Hochberg method (Benjamini & Hochberg, 1995). Sites in individual samples were considered differentially methylated if they had a *p*-value below 0.01 and a median methylation difference of at least 50% relative to the sites in all of the samples they were compared to.

Assessing the biological impact of genetic and epigenetic mutations

The biological impact of SNPs and small indels was assessed using VEP (McLaren *et al*., 2016) (version 102). Genes were obtained from the TAIR10 genomic annotation of Ensembl Plants (release version 40) and from the annotation of the improved *A. thaliana* mitochondrial assembly (BK010421.1). In addition, we intersected genetic and epigenetic variants with protein-coding genes, the 500 bp upstream of such genes, and transposable elements annotated in the *A. thaliana* genome using bedtools (Quinlan & Hall, 2010) (version 2.27.1) intersect, keeping hits with an overlap of at least 1 bp.

Bulked segregant analysis (BSA)

F_2 progeny obtained from an F_1 generated from a cross between Col-0 (mother) and a zinc tolerant ZSI2 (father) plant were grown under moderate zinc stress until several inflorescences had formed. Plants were then grouped into two phenotypic categories, tolerant or sensitive, for their response to the high zinc concentration based on their growth and visual signs of stress (i.e. chlorosis). Flower heads were sampled and pooled of approximately 70 plants a per phenotypic group. DNA was isolated using the same CTAB method as described before.

Library preparation was done using the Hackflex protocol (Gaio *et al*., 2022). Samples were pooled, and the fraction showing reads between 300 and 500 bp in size were selected and sequenced for on average 40X whole genome coverage sequencing with Novogene (UK) Ltd. The SNP and indel calling workflow consists of four steps: (1) read trimming, (2) read alignment, and (3) variant calling. Step 1: Reads were trimmed using Cutadapt (Martin, 2011) (version 1.18). This step clipped sequences that matched at least 90% of the total length of one of the adapter sequences provided in the NEBNext Multiplex Oligos for Illumina (Index Primers Set 1). In addition, it trimmed bases from the 5′ and 3′ ends of reads if they had a Phred score of 20 or lower. Reads shorter than 70 bp after trimming were discarded. Step 2: Trimmed reads were aligned to a modified version of the *A. thaliana* Col-0 reference genome (TAIR10, European Nucleotide Accession number: GCA_000001735.2) which contained an improved assembly of the mitochondrial sequence (Sequence Read Archive accession number: BK010421) (Sloan *et al*., 2018) using bwa mem (version 0.7.10-r789) (H. Li, 2013) with default parameters. The resulting alignment files were sorted and indexed using samtools (version 1.3.1) (H. Li *et al*., 2009) Alignment files of libraries generated from the same accession were merged using Picard MarkDuplicates (https://broadinstitute.github.io/picard/), called through the GATK suite (version 4.0.2.1) (McKenna *et al*., 2010). Picard MarkDuplicates was also used to mark duplicate read pairs, using an optical duplicate pixel distance of 2500, which is appropriate when working with patterned Illumina flow cells. Step 3: SNPs and indels were called by running FreeBayes (Garrison & Marth, 2012) (version 1.3.1-dirty) with alignment files of all samples as input, using default parameters. Step 4: the two most abundant alleles of a variant taken and variants with a quality score below 1 were excluded using VcfFilter. Step 5: the VariantsToTable function from the GATK suite was used to generate a tabular format that can be used in subsequent bulked segregant analyses.

Bulked segregant analysis was done using QTLseqr (Mansfeld & Grumet, 2018) (version 0.7.5.2) for all possible pairs between the four phenotypic classes. Variants that occurred less than 20% in both pools and extremely low and high coverage SNPs, with a minimum of 30 and maximum of 100, were excluded. A sliding window size of 300 kb was used. Other steps were used in default mode, and as a false discovery rate q = 0.01 was used.

Supplementary Figures

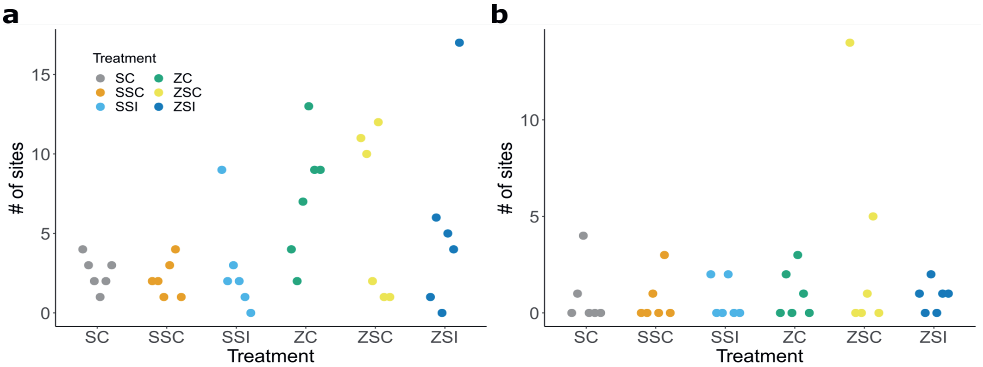


Figure S1: **Differentially methylated CHG and CHH sites in plans of the fifth generation grown under different treatments.** (a) The number of differentially methylated CHG sites in individual plants of each treatment. (b) The number of differentially methylated CHH sites in individual plants of each treatment.

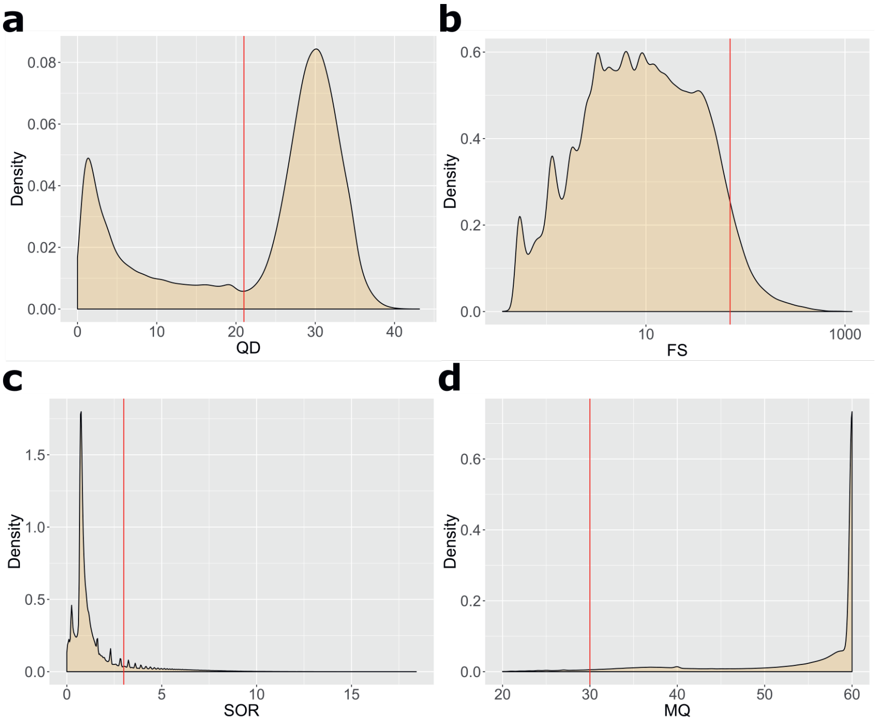


Figure S2: **Quality control of SNPs and small indels.** (a-d) Density plots of the variant annotations quality normalized by read depth (QD) (a), Fisher's exact test of strand bias (FS) (b), symmetric odds ratio test of strand bias (SOR) (c), and mapping quality (MQ) (d) of nuclear SNPs and small indels. Thresholds used to filter putative false positives are shown in red. Similar plots were used to filter mitochondrial and chloroplast variants.

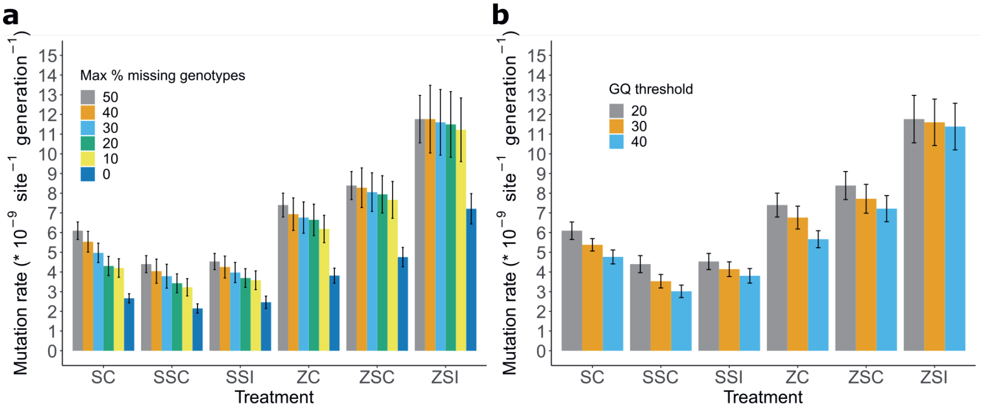


Figure S3: **Mean SNP mutation rates at different thresholds for genotype quality (GQ) and the percentage of missing genotypes.** Although absolute rates change between different minimum GQ (a) and maximum allowed percentages of missing genotypes (b), relative differences between mutation rates are retained. Error bars depict standard errors of the mean. The same trends were seen when computing mutation rates of small insertions and deletions.

Chapter 4

Local adaptation of *Arabidopsis thaliana* in a small geographic region with mild environmental clines

Raúl Y. Wijfjes[1,6,*], René Boesten[2*], Frank F. M. Becker[2], Tom P. J. M. Theeuwen[2], Basten Snoek[3], Maria Mastoraki[2], Jelle J. Verheijen[2], Nuri Güvencli[2], Lissy-Anne M. Denkers[4], Maarten Koornneef[2], Fred van Eeuwijk[5], Sandra Smit[1], Dick de Ridder[1,†], and Mark G.M. Aarts[2,†]

[1] Bioinformatics Group, Wageningen University & Research
[2] Laboratory of Genetics, Wageningen University & Research
[3] Theoretical Biology and Bioinformatics, Utrecht University
[4] Department of Plant Physiology, Green Life Science Research Cluster, Swammerdam Institute for Life Sciences, University of Amsterdam
[5] Biometris, Wageningen University & Research
[6] Current affiliation: Faculty of Biology, Ludwig Maximilian University of Munich
[*] These authors contributed equally to this work.

Supplementary methods and tables are available on bioRxiv:
https://doi.org/10.1101/2023.09.18.558200

Abstract

Natural populations of *Arabidopsis thaliana* provide powerful systems to study adaptation of wild plant species. Previous research has predominantly focused on global populations or accessions collected from regions with diverse climates. However, little is known about the genetics underlying adaptation in regions with mild environmental clines. We have examined a diversity panel consisting of 192 *A. thaliana* accessions collected from the Netherlands, a region with limited climatic variation. Despite the relatively uniform climate, we identified compelling evidence of local adaptation within this population. Notably, semidwarf accessions occur at a relatively high frequency near the coast and these displayed enhanced tolerance to high wind velocities. Additionally, we evaluated the performance of the population under iron deficiency conditions and found that allelic variation in the *FSD3* gene affects tolerance to low iron levels. Moreover, we explored patterns of local adaptation to environmental clines in temperature and precipitation, observing that allelic variation at *LARP1c* likely affects drought tolerance. Not only is the genetic variation observed in a diversity panel of *A. thaliana* collected in a region with mild environmental clines comparable to that in collections sampled over larger geographic ranges, it is also sufficiently rich to elucidate the genetic and environmental factors underlying natural plant adaptation.

Introduction

Adaptation is defined as the process through which a population attains higher fitness in its native environment than non-native populations sampled from different sites (Kawecki & Ebert, 2004). Unraveling the genomic and physiological basis of adaptation in plants is a central question in modern plant biology. From an evolutionary perspective, linking plant genotypes to adaptive traits helps to clarify how environmental gradients drive natural selection of both standing and novel variation, which ultimately may result in speciation (Sobel *et al*., 2010). It also enables us to model whether natural populations are able to adapt to future environmental conditions (Bay *et al*., 2017; Exposito-Alonso *et al*., 2019) and provides insights on processes or even genes to focus on when breeding crops for enhanced abiotic stress tolerance (Huang & Han, 2014).

The model plant species *Arabidopsis thaliana* provides an excellent system to study natural plant adaptation (Weigel & Nordborg, 2015). Recent studies of this species have shifted from the global HapMap population (Li *et al*., 2010), traditionally used for genome-wide association studies (GWAS), to populations sampled from regions containing a range of contrasting climates, including Sweden (Long *et al*., 2013), the Iberian Peninsula (Tabas-Madrid *et al*., 2018), and the south-west of France (Frachon *et al*., 2018), as these local populations are expected to minimize genetic heterogeneity and thus provide more statistical power for detecting quantitative trait loci (QTLs) (Brachi *et al*., 2013; Korte & Farlow, 2013). While studies on such regional populations have successfully identified several adaptive loci (Frachon *et al*., 2018; Tabas-Madrid *et al*., 2018), we know comparatively little about the loci driving adaptation in regions having a milder diversity in environmental conditions.

The Netherlands, a region which has remained largely under-sampled so far, provides an excellent opportunity to address this issue. This region covers an area comparable in size to the south-west of France, but much smaller than Sweden and the Iberian Peninsula. Compared to the other three regions, the Netherlands has only mild climatic clines (https://www.knmi.nl/klimaat-viewer), but past work indicated that local adaptation of *A. thaliana* may still be expected (Barboza *et al*., 2013). The Netherlands contains a moderate frequency of accessions with short inflorescences, mediated by the same loss-of-function allele of the *GA5* gene that managed to spread over more than 100 kilometres in the western part of the country (Barboza *et al*., 2013). There is reason to believe that more such signatures of local adaptation can be found, for instance based on the small differences in average temperature and annual precipitation, or the soil type on which *A. thaliana* grows in the Netherlands (De Vries *et al*., 2003). While it can be found on the clay and peat regions in the west and north of the country, it mainly grows on sandy soils, which largely occur in the east and south, and in the coastal dunes. It is often found in roadsides and gardens in which substantial movement of soils and seeds from other regions has occurred in the past.

To comprehensively address the contribution of genetic variation to plant adaptation in the Netherlands, we generated the Dutch

Arabidopsis thaliana Map (DartMap) panel, a collection of 192 *A. thaliana* accessions sampled from different sites in the country. We show that this panel contains a surprisingly high level of genetic diversity and provide evidence of local adaptation to mild climatic clines. Moreover, we present strong evidence that variation at two loci is involved in mediating adaptation to wind tolerance and iron deficiency. Our work demonstrates that plant populations sampled in small geographic regions with a low diversity in environmental conditions can contain a level of genetic variation that is sufficiently rich to facilitate local adaptation.

Results

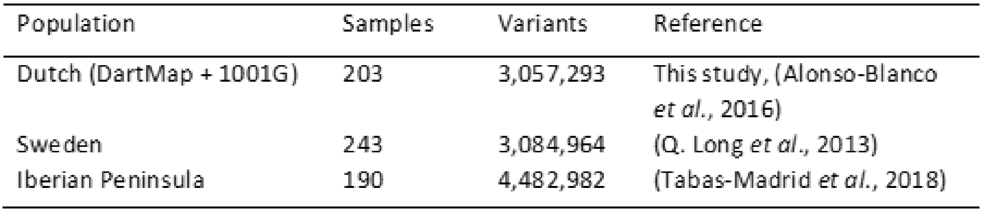
The degree of genetic diversity in the Netherlands is typical of that of most European regions

The global *A. thaliana* collection of the 1001 Genomes (1001G) Project (Alonso-Blanco *et al*., 2016) includes only 11 lines from the Netherlands, which likely represents only a small portion of the region's genetic variation due to limited dispersal of minor alleles (Tabas-Madrid *et al*., 2018). Therefore, we used the DartMap panel and the Dutch 1001G accessions to characterize the overall degree of genetic diversity of *A. thaliana* in the Netherlands. In total, we identified 2,712,612 SNPs and 353,974 indels in the DartMap panel and the Dutch 1001G accessions (collectively referred to as DartMap + 1001G panel for brevity) relative to the *A. thaliana* Col-0 nuclear genome reference sequence. Moreover, we detected 487 SNPs and 209 indels in the chloroplast sequence, and 137 SNPs and 15 indels in the mitochondrial sequence.

Furthermore, we examined copy number variation in the DartMap panel, excluding the Dutch 1001G accessions because they were sequenced with a different sequencing platform. We detected 29,155 copy number variants (CNVs) relative to the Col-0 reference genome. Most of these are less than 500 bp (Figure S1a) and are deletions (Figure S1b), which is in line with the distribution of structural variants found in an earlier study on the full 1001G collection (Göktay *et al*., 2021). We found that 448 genes overlap with CNVs predicted to disrupt gene function and present at moderate frequencies within the DartMap panel (Table S1) (considering deletions only, as we could confirm that these were called with high precision (Figure S2)). The 448 genes are significantly enriched for genes involved in disease resistance ($P = 4.39\times10^{-6}$).

To contextualize the degree of genetic diversity of the DartMap + 1001G panel, we compared it to similarly sized collections from Sweden (Long *et al*., 2013) and the Iberian peninsula (Tabas-Madrid *et al*., 2018), using the same SNP and indel calling approach. Considering only SNPs and indels that are polymorphic in the DartMap + 1001G panel (i.e. at least one accession contains a reference allele), the Swedish collection contains a similar number of variants, while the Iberian one is more diverse (Table 1). This disparity can be attributed to the presence of “relict” accessions in the Iberian collection, which originated from *A. thaliana* populations in Europe before the last ice age and are genetically distinct from the more recent “non-relict” group that repopulated Europe thereafter (Alonso-Blanco *et al*., 2016).

Table 1: **The number of polymorphic bi-allelic SNPs and indels ("Variants") found in regional *A. thaliana* populations.**



Population	Samples	Variants	Reference
Dutch (DartMap + 1001G)	203	3,057,293	This study, (Alonso-Blanco *et al.*, 2016)
Sweden	243	3,084,964	(Q. Long *et al.*, 2013)
Iberian Peninsula	190	4,482,982	(Tabas-Madrid *et al.*, 2018)

Most SNPs and indels detected in the DartMap + 1001G panel are shared among multiple accessions and the distribution of their minor allele frequencies is similar to that of the Swedish collection (Figure 1a). In contrast, the Iberian collection contains a relatively larger fraction of rare variants (Figure 1a), possibly because the mountain systems in this region strongly limit the spread of minor alleles (Tabas-Madrid *et al.*, 2018).

To assess the genetic diversity of the DartMap + 1001G panel, we calculated pairwise genetic distances between individual accessions. Most pairs of accessions differ at approximately 320,000-340,000 variant sites, similar to the Swedish collection and most pairs of the Iberian collection (Figure 1b). However, the Iberian collection shows a right-skewed distribution of pairwise genetic distances, indicating the presence of accessions that are genetically more distant from each other compared to the majority (Figure 1b). This tail in the distribution aligns with previous findings of genetic distance between Iberian accessions and attributed to the presence of the relict accessions in this region (Alonso-Blanco *et al.*, 2016). Taken together, our analyses indicate that genomic variation of *A. thaliana* in the Netherlands predominantly originated from the post-glacial expansion event from which most *A. thaliana* accessions in Europe descended (C.-R. Lee *et al.*, 2017).

Genetic diversity of *A. thaliana* in the Netherlands was shaped by ancient and contemporary forces

We explored the demographic and evolutionary history of the DartMap + 1001G panel using population structure analyses based on the genetic distance between individual accessions, which have proven useful for this purpose in previous studies of *A. thaliana* (Alonso-Blanco *et al.*, 2016; Tabas-Madrid *et al.*, 2018). We computed population structure of the DartMap + 1001G panel through hierarchical clustering based on pairwise organellar (Figure 2a) and nuclear (Figure 2b) genome-wide similarity, including all accessions of the 1001G collection sampled in Belgium, France, Germany, and the United Kingdom to evaluate relatedness between Dutch *A. thaliana* accessions and those of nearby countries. The dendrograms based on chloroplast and mitochondrial variation consist of three nearly equidistant groups (Figure S3) with identical spatial distributions across the Netherlands (Figure 2a), as expected for organellar sequences that are both predominantly maternally inherited.

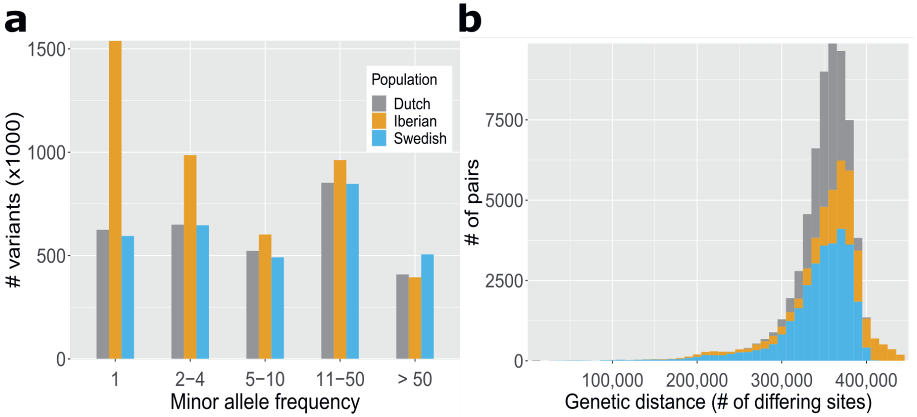


Figure 1: **Comparing the genetic diversity of three regional populations.** (a) Minor allele frequency of bi-allelic SNPs and indels. (b) Genome-wide similarity between pairs of samples (bars are overlapping).

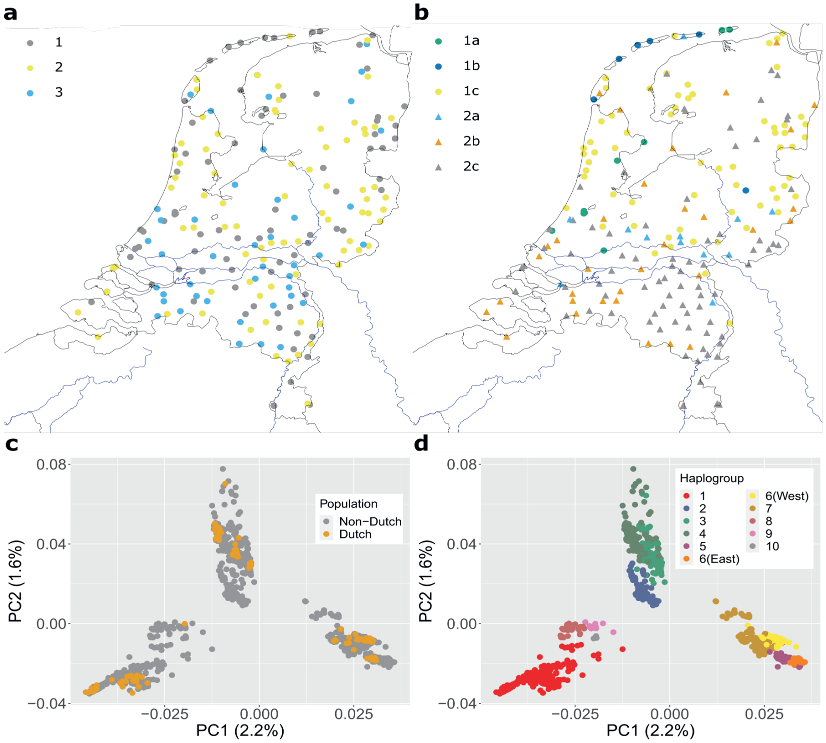


Figure 2: **Population structure of *A. thaliana* in the Netherlands.** (a-b) Geographical location of Dutch accessions clustered based on organellar (a) and nuclear (b) genome-wide similarity. Colours are used to distinguish accessions of different clusters. The main rivers of the Netherlands are depicted by blue lines. (c) Principal component analysis of chloroplast variants in Dutch accessions and the global collection of *A. thaliana* analysed in a previous study (Hsu et al., 2019). (d) Samples in (c) coloured according to the chloroplast haplogroups reported in a previous study (Hsu et al., 2019).

A previous study showed that European accessions contain several highly diverged chloroplast haplogroups of which the estimated time of divergence considerably predates the post-glacial expansion event (Hsu *et al*., 2019). A principal component analysis based on chloroplast variation between individual accessions of the DartMap + 1001G panel and of a global collection of *A. thaliana* separate Dutch accessions in three different groups (Figure 2c) that correspond to different subsets of chloroplast haplogroups identified in that same study (Figure 2d). This implies that the distinct organellar groups of the DartMap + 1001G panel represent ancient variation, while the nuclear groups reflect more recent genomic differentiation, explaining the limited overlap in spatial distribution between the two (Figure 2a and b).

The spatial distribution of the nuclear groups provides insights into the factors that shaped genomic variation in the DartMap + 1001G panel following the postglacial expansion event. Based on nuclear variation, the panel can be divided into two main groups, predominantly separated by the central Rhine and Waal rivers (group 1 to the north and group 2 to the south) (Figure 2b). The rivers likely acted as natural dispersion barriers (Pico *et al*., 2008), restricting genetic flow between the two groups. It is worth noting that *A. thaliana* is generally not found in clay soil accompanying the rivers, except in residential areas where it has been introduced through human-induced transport of soil from other regions. Group 1 can be further divided into three subgroups (Figure 2b), of which two contain accessions that are closely related to each other (subgroups 1a and 1b) and less related to the rest of the population (Figure S4). Subgroup 1a contains nine accessions with short inflorescences, referred to as “semidwarfs”, mainly concentrated in the western part of the Netherlands (Figure 2b). Subgroup 1b is predominantly found on the Dutch islands in the north, implying that these islands were colonized by a limited number of founder accessions, mostly from group 1. Group 2 consists of accessions that are genetically similar to accessions from Germany (subgroup 2a) respectively France and the United Kingdom (subgroup 2b) (Figure S5), suggesting that these are part of a large genetic group spread over Northwestern Europe.

Besides large clusters, we identified 27 pairs of accessions that are near duplicates in terms of nuclear similarity (Table S2), with differences of fewer than 50,000 sites between them. This level of similarity is significantly lower than the median pairwise similarity of 354,388 sites (Figure 1a). Strikingly, they include 8 out of 9 pairs of the semidwarf accessions of group 1a, which are separated by more than 100 km (Table S2). The widespread geographical distribution of highly identical genotypes that all have the same semidwarf phenotype suggests there may be a selective advantage to this trait, similar to the near-identical and widespread atrazine-resistant genotypes distributed along the United Kingdom railway system, which we identified previously (Flood, Van Heerwaarden, *et al*., 2016).

Semidwarf *ga5* mutants display tolerance to windy conditions

A previous study indicated a higher frequency of semidwarfs in the Netherlands compared to the global estimated frequency of at least 1% (Barboza *et al*., 2013). In our analysis, we identified ten semidwarf accessions in the DartMap panel, three of which overlapped with those from the previous study(Barboza *et al*., 2013). Almost all examined semidwarf accessions are loss-of-function mutants at the *GA5* locus, which encodes the *GIBBERELLIN 20-OXIDASE 1* (*GA20OX1*) gene (Barboza *et al*., 2013).Further investigation of the genomic sequences of *GA5* revealed that eight semidwarf accessions are genotypically nearly identical and share a previously described *ga5* splice site mutation specific to Dutch semidwarfs (Barboza *et al*., 2013). Of the remaining two accessions, one contains an undescribed 52 bp deletion in the last exon of *GA5*, while the other contains no mutations in the *GA5* coding region.

The eight near-identical semidwarfs that share the same splice site mutation are part of nuclear subgroup 1a (Figure 2b), whereas the semidwarf accession without *GA5* coding sequence mutations is genetically distinct and part of subgroup 1b (Figure 2b). Subgroup 1b consists of five accessions found exclusively on the Dutch islands in the North, with four of them displaying relatively short inflorescence lengths (among the top 15% shortest inflorescence length in the DartMap). These four accessions share a unique 25 bp insertion at 557 bp upstream of the *GA5* transcription start, while the fifth accession in this subgroup has an average inflorescence length and lacks this mutation. Both subgroups (1a and 1b) are genetically most related to each other and both are found in the West of the Netherlands (Figure 2b). The appearance of two distinct *ga5* mutations in accessions with short inflorescences dispersed across the Netherlands suggests that semidwarfism provides an adaptive advantage.

Dwarfism in plants is a common adaptation to high altitude (Billings & Mooney, 1968; Grace, 1988; Luo *et al*., 2015), although it remains unclear which aspect of high altitude causes this. The majority of the Dutch semidwarfs are collected within 15 km of the coast, often only a few meters above sea level, so there is no association with high altitude. However, these semidwarfs share a characteristic with their original collection locations, namely the presence of high wind speeds, particularly between December and March (Figure 3a). Given that high wind speeds are also a characteristic of high-altitude environments, we aimed to investigate whether the presence of high wind speeds could potentially create a selective environment favoring semidwarf *A. thaliana*.

We grew six Dutch semidwarf accessions and five tall accessions, in a climate chamber equipped to provide controlled wind speeds typically experienced in coastal areas of the Netherlands in March/April. Moreover, we included Col-0, Ler and the ga5 loss-of-function mutants in Col-0 and Ler backgrounds for a direct comparison of near isogenic lines differing only at the *GA5* locus. Despite our preselection for similar flowering time (Table S3), based on greenhouse data, the semidwarfs flowered significantly later than the taller accessions (Figure 3b), and the wind treatment therefore affected the semidwarfs for a shorter period while bolting and flowering than the tall accessions. This likely had little effect on our final results though, as all accessions were exposed to windy conditions for the same time and all plants started to flower before the end of the experiment. The tall accessions are strongly affected by the wind, as evidenced by a significant reduction in stem length (approximately 30% on average) and fresh weight (approximately 36% on average) (Figure 3c and 3d). In contrast, the semidwarfs also showed a significant reduction in stem length, but the average reduction was only around 10% and no significant decrease in fresh weight was observed (Figure 3d). Similar effects were observed for Col-0 and Ler, where the wind treatment significantly decreased their stem length and fresh weight, whereas the respective *ga5* loss-of-function mutants remained unaffected (Figure 3c and d). Collectively, these results strongly suggest that the presence of the *ga5* mutation contributes to wind tolerance.

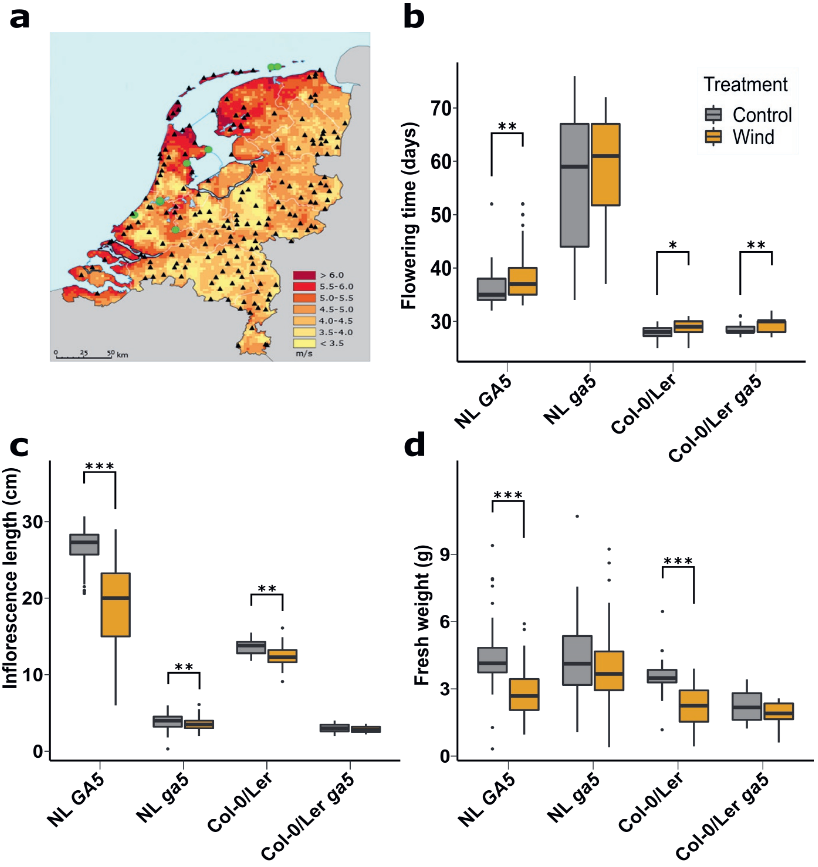


Figure 3: **Phenotypic response to windy conditions.** (a) Average annual wind speed (between 1981-2010) in the Netherlands as measured at ground level. Green circles indicate semidwarf accessions that share the same splice site mutation, black triangles indicate all other DartMap accessions. Figure adapted from https://www.knmi.nl/klimaat-viewer/kaarten/wind/gemiddelde-windsnelheid/jaar/Periode_1981-2010. (b-d) Plant responses to absence or presence of windy conditions (grey and orange respectively) for (b) flowering time, (c) inflorescence length one week after the first flower opened, (d) fresh weight at flowering $^{*}P$ < 0.05; $^{**}P$ < 0.01; $^{***}P$ < 0.001, two-tailed Student's t-test.

Flowering time variants display a latitudinal cline

The transition from vegetative to reproductive stage is a key life history trait in *A. thaliana*, influenced by seasonal cues like day length and temperature(Andrés & Coupland, 2012). Previous GWA analyses of flowering time yielded few significant associations, mainly due to allelic heterogeneity (Atwell *et al*., 2010; Tabas-Madrid *et al*., 2018; L. Zhang & Jiménez-Gómez, 2020). We expect this should be less of a problem in a regional population with more closely related individuals. Therefore, we performed GWA analysis of flowering time, with and without vernalization, using the DartMap panel. We discovered a highly significant association between flowering time and a 16 bp indel near the end of the first exon of the *FRIGIDA* (*FRI*) gene (position 269960 on chromosome 4 of Col-0), irrespective of vernalization treatment (Figure 4a). This indel is known to cause a premature stop codon in the *FRI* coding sequence, leading to a non-functional *FRI* allele (Andrés & Coupland, 2012; Johanson *et al*., 2000; Koornneef *et al*., 1994).

FRI is a key regulator of flowering time in *A. thaliana* and loss-of-function alleles are widespread in nature (L. Zhang & Jiménez-Gómez, 2020). As the variant we picked up by GWA analysis is unlikely to be the only loss-of-function mutation to occur in the Netherlands, we examined all polymorphisms in the *FRI* coding sequences in the DartMap and found a total of 41 *FRI* alleles. When compared to the *FRI-H51* allele, that encodes a fully functional protein and is regarded as the ancestral *FRI* allele based on sequence comparisons with *A. lyrata* (Le Corre *et al*., 2002; L. Zhang & Jiménez-Gómez, 2020), another 11 loss-of-function *fri* alleles are identified across 88 accessions. Nearly all of the accessions with a loss-of-function allele are early flowering (Figure 4b). Although the Netherlands has a small latitudinal range, we examined if there is a latitudinal cline in the diversity panel regarding flowering time, consistent with previous findings in Northern European and Mediterranean accessions (Stinchcombe *et al*., 2004). While we did not observe such a cline when considering the entire population, excluding accessions with *fri* loss-of-function alleles revealed a latitudinal cline (Figure 4c), suggesting the involvement of other loci, apart from *FRI*, in controlling flowering time. These analyses collectively demonstrate the suitability of the DartMap panel for conducting GWA analyses on phenotypic variation relevant to local adaptation.

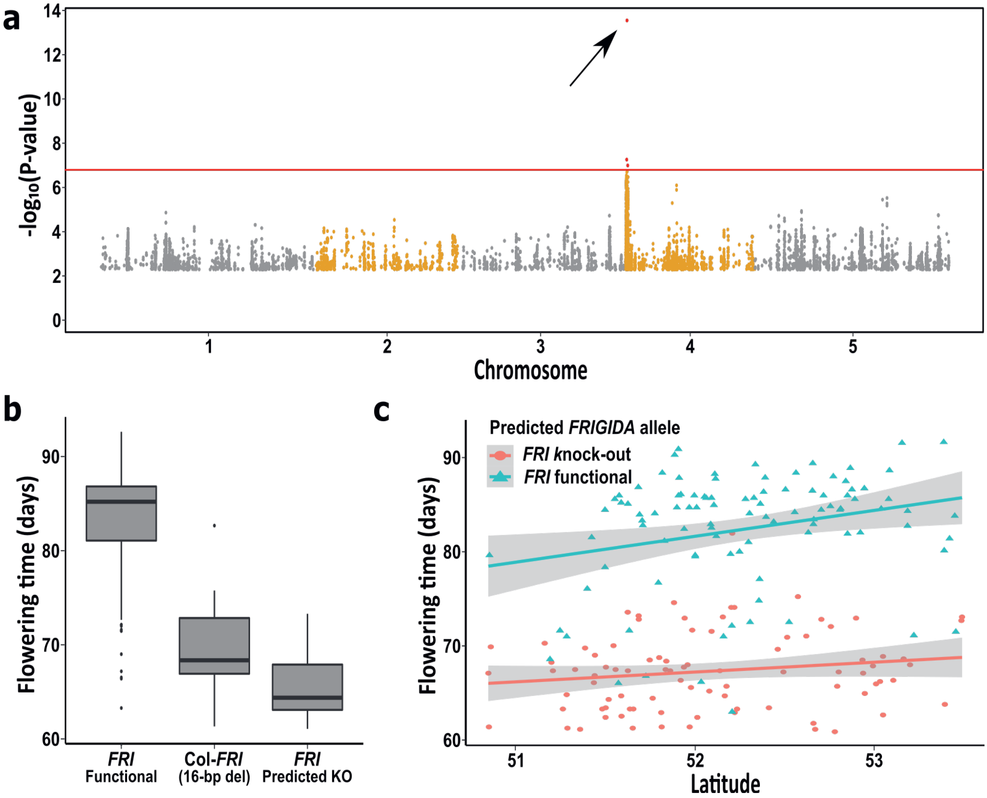


Figure 4: **GWA analysis of flowering time in the DartMap panel upon vernalization.** (a) Manhattan plot for days until flowering after a six-week vernalization treatment, identifying one highly associated polymorphism located in the *FRIGIDA* gene (arrow). The red line corresponds to a permutation-based threshold level of significance at $P < 0.05$. (b) Flowering time distribution of three *FRIGIDA* variants in the DartMap panel. The Col-0 variant corresponds to a 16 bp deletion causing a premature stop codon in the *FRIGIDA* coding sequence. (c) Average days until flowering for DartMap accessions relative to their latitudinal site of origin with linear fit for accessions with a predicted loss-of-function fri allele (red, $R^2 = 0.016$, $P = 0.128$) or a functional *FRI* allele (blue, $R^2 = 0.082$, $P = 0.002$).

Local adaptation to the Dutch climate

The Dutch climate is characterized as a temperate maritime climate, which is relatively uniform across the Netherlands. However, there are mild and gradual clines in environmental variables, with slightly higher temperatures and less precipitation in inland areas compared to coastal areas. One advantage of having a densely and uniformly sampled population is the potential to detect whether local adaptation to these gradual environmental clines has influenced genetic diversity. To test this hypothesis, we conducted a genome-environment association (GEA) analysis on 19 quantitative climatic variables related to temperature and precipitation (Table S4 and Figure 5a as example). We identified multiple QTLs for these variables, but the most notable QTL is located on Chr. 4 (Figure 5b), which we named *qDPT*

(QTL for the Dutch Precipitation and Temperature climatic variables). This QTL is associated with seven climatic variables (Table S4).

Further examination of *qDPT* reveals that it appears to consist of two separate QTLs, approximately 230,000 base pairs apart (Figure S6a). The SNPs that are found most frequently with the highest association in each QTL are SNP $4^{16770900}$ and SNP $4^{17000635}$ with minor allele frequencies (MAF) of 14.2% and 17.5%, respectively. To determine whether the -log(*P*) peaks represent distinct QTLs or a single one, we calculated pairwise r^2 values between SNP $4^{16770900}$ and all other biallelic SNPs in the selected genomic region (16,700,000–17,200,000 bp). This analysis revealed only weak evidence of linkage disequilibrium (LD) (pairwise $r^2 = 0.52$) between the two SNPs most strongly associated with the trait (Figure S6b). Hence, we consider these to be two separate QTLs. For convenience, we will refer to them as '*qDPT-A*' and '*qDPT-B*', with their corresponding SNPs labelled as *A*/*a* for SNP $4^{16770900}$ and *B*/*b* for SNP $4^{17000635}$.

The GEA analyses provide insights into whether specific alleles are correlated with particular climatic variables. A strong association at a specific locus can indicate that allelic variation at that locus is important for local adaptation (Ferrero-Serrano & Assmann, 2019). However, it is important to note that a strong correlation could also arise if the climatic variable aligns with the population structure, potentially resulting in a statistical artifact. We confirmed that this is not the case. First of all, kinship correction is used to perform the GEA analyses, and this should drastically reduce the chance of finding loci that strongly link to the population structure. Secondly, we observe that the geographical distribution of the population structure appears to be mostly affected by the river systems (Figure 2b), which can be regarded as a South – North cline. The climatic variables that map at *qDPT* follow a mostly East – West cline. Moreover, the minor allele of *qDPT-A* is found in five out of six nuclear subclusters (the common allele is found in all subclusters), and for *qDPT-B* both alleles are found all subclusters. In conclusion, identification of these QTLs is not an artifact due to the general population structure and is likely to represent a locus associated with plant adaptation to the environment.

It is not immediately apparent to which specific climatic variable(s) this locus contributes. In the Netherlands, the climate is primarily influenced by the North Sea, which affects various environmental factors such as temperature, wind speed and precipitation patterns and variability. Therefore, it is not surprising that the same peak is consistently observed, given the high correlation between climatic variables associated with *qDPT*. Consequently, unravelling the particular climatic variable or combination of variables to which this specific locus may be adaptive is challenging. However, it is worth noting that the climatic variables associated to *qDPT* are all connected to seasonality, while annual means in precipitation and temperature are not associated. This suggests that the adaptation may be related to conditions that prevail during certain parts of the year, such as one or two seasons, or to adaptations that enable the plant to effectively tolerate greater variability. To explore which climatic factor may be involved in this regard, we tested if allelic variation at *qDPT* resulted in a differential response to a gradual decrease in water availability.

As *qDPT-A* and *qDPT-B* are in close proximity, we selected a set of accessions with combinations of alleles for these QTLs (13 accessions with '*AB*', 7 with '*Ab*', 1 with '*aB*' and 7 with '*ab*'; Table S5) aiming to disentangle the individual contributions of the QTLs. Initially, all accessions were grown under well-watered conditions. Subsequently, after two weeks, half of the replicates for each accession were no longer watered for eleven days. We regularly measured the response in projected leaf area (PLA) during this period. Due to an uneven distribution of accessions across the genotypic combinations for these QTLs, we compared accessions with '*Ab*' genotypes against those with '*ab*' genotypes (7 accessions per group) to assess the effect of *qDPT-A*. For *qDPT-B*, all 28 accessions (14 accessions per *qDPT-B/b* allele) were included in the analysis. Our results indicate a significant effect of *qDPT-B* on the PLA response to drought (two-way ANOVA with *qDPT-B/b* allele and days post drought treatment as variables; $P = 0.003$) (Figure 5c and 5d). However, we did not observe a significant effect for *qDPT-A*. Subsequently, we examined the distribution of *qDPT-B* alleles across the Netherlands and the 1001G panel. We found that SNP $4^{17000635}$ (*qDPT-B*) has a much higher MAF in the Netherlands (MAF = 17.5%) (Figure 5e) compared to 1001G panel (MAF = 3.1%) (Figure 5f). The association scores at *qDPT-B* varied considerably depending on the specific climatic factor being analysed.

We propose that the region encompassing the genomic locus with a LOD > 6, which is consistently observed across different mappings, is the most likely location for the causal gene(s). This region comprises 13 genes spanning from *At4g35820* to *At4g35940* (Table S6). Within this region, there are a few genetic variants that are in linkage ($r^2 > 0.4$) with SNP $4^{17000635}$ and potentially impact the protein. Particularly noteworthy are two non-synonymous variants in *OSCA4.1* (*At4g35870*), also known as early-responsive to dehydration stress protein (*ERD4*). The first variant constitutes a change from methionine (neutral-polar) to isoleucine (hydrophobic) at the 10th position, while the second variant changes an isoleucine to leucine (both hydrophobic) at the 641st position. Presently, it is unknown whether these changes affect functionality of this gene. We tested T-DNA insertion lines for 11 of these candidate genes in response to control and drought treatments and measured Fv/Fm as a proxy of plant stress (Maxwell & Johnson, 2000) (Figure 5g). For the remaining two genes, *At4g35930* and *At4g35940*, no T-DNA insertion lines were available. The mutant of *LARP1c* (*At4g35890*) has significantly lower Fv/Fm values in the drought treatment compared to Col-0, while it does not differ from Col-0 in the control treatment (Figure 5g). Additionally, the mutant performance does not differ significantly from Col-0 in PLA under the control treatment, but is significantly impaired under the drought treatment (Figure 5h). These results pinpoint *LARP1c* as the most prominent candidate gene to underly the difference in response to drought.

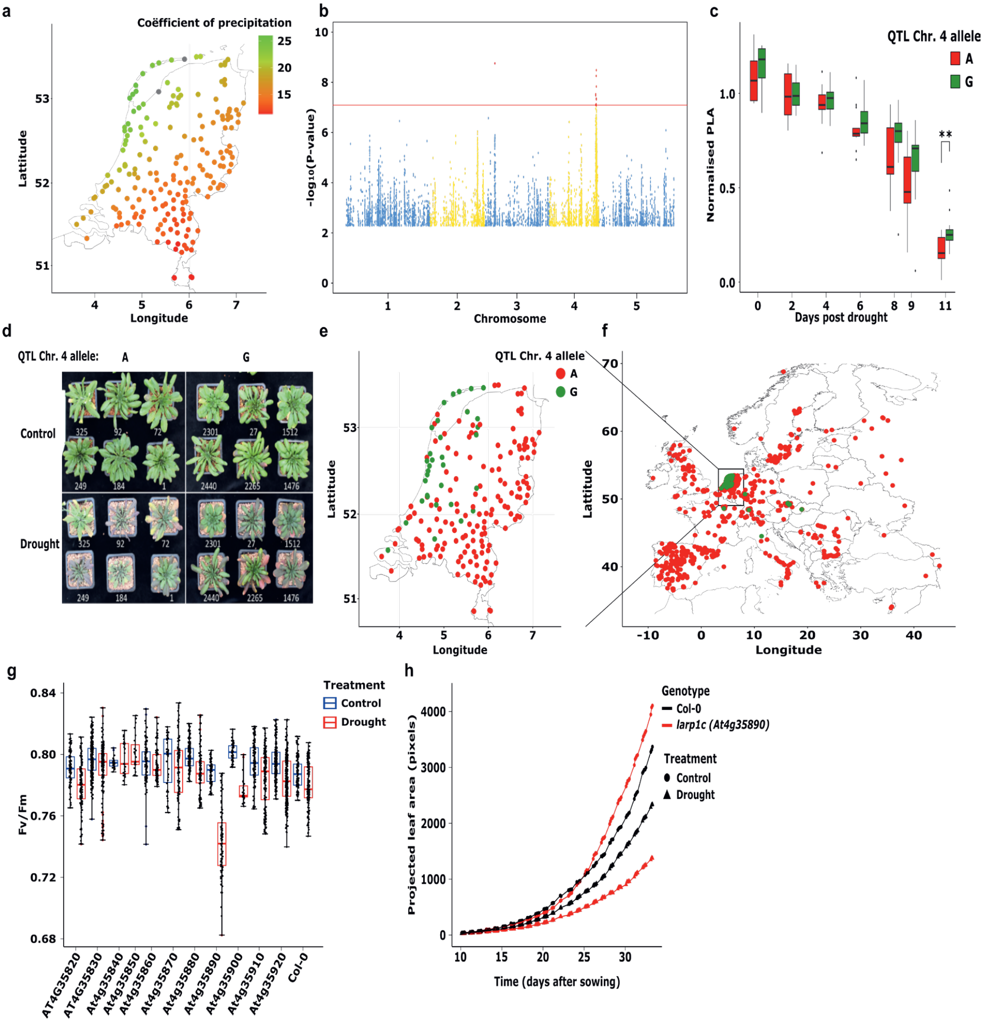


Figure 5: **Genome-environment association (GEA) analysis on climatic variables related to precipitation and temperature.** (a) Geographical distribution of precipitation seasonality (coefficient of variation) expressed in Seasonality Index (SI) units for the DartMap. SI values < 0.19 are defined as 'precipitation throughout the year', whereas values between 0.20 and 0.39 are defined as 'precipitation throughout the year, but with a definite wetter season'. (b) Manhattan plot for the GEA related to precipitation seasonality (coefficient of variation). (c) Response to drought treatment for accessions differing at their *qDPT-B* allele. Boxplots represent the normalised projected leaf area (PLA) calculated per accession by dividing the PLA in the drought treatment over the PLA in the control treatment for a set of 14 randomly chosen accessions per allele (depicted as an adenine (A) or guanine (G) nucleotide at SNP $4^{17000635}$). (d) Pictures of six representative accessions per *qDPT-B* allele in the control and drought treatment, taken 11 days after the start of the drought treatment. (e) Geographical distribution of the two alleles ('A' and 'G') across the Netherlands and (f) Europe. (g)) Fv/Fm as a proxy of plant stress of Col-0 and candidate gene T-DNA insertion lines. (f) The response of PLA of Col-0 and the *larp1c* mutant.

Natural allelic variation for *FSD3* affects iron deficiency tolerance

Although the Netherlands is relatively uniform in climate, the soil type composition differs substantially across the country, with clay, peat and sand forming the major types (Hartemink & Sonneveld, 2013). This could be an important driver for local adaptation in the Netherlands, as nutrient composition and particularly nutrient availability (Veer, 2006), differs between soil types. Iron (Fe) in particular is known to vary in availability and shortage of Fe will severely limit plant growth (Hell & Stephan, 2003). We determined the light-use efficiency of photosystem II (PSII) electron transport (ΦPSII) of the DartMap diversity panel grown under control and Fe-deficient conditions, through chlorophyll fluorescence measurements. Low ΦPSII can serve as an early indicator of Fe deficiency (Terry, 1980), as Fe is an essential component of ironsulphur, and heme proteins that play pivotal roles in photosynthesis and respiration(Hänsch & Mendel, 2009).

The low Fe supply leads to only mild visible symptoms, such as reduced plant growth and mild leaf chlorosis, but causes extensive variation for ΦPSII, much more than in the control condition (Figure 6a). GWA mapping identified a single associated region, on chromosome 5, with the most significantly associated SNP located at position 7,856,132 (Figure 6b). The same locus, and the same SNP, is also found when the average ΦPSII values in the Fe-deficient condition, and the ratio between ΦPSII in Fe deficiency versus control, are used as traits for GWA mapping (Figure S7). To identify candidate genes for the causal allelic variant, we limited the associated region to 22 kb, based on the genome positions of SNPs that are in linkage disequilibrium (LD) ($r^2 > 0.55$) with the most significantly associated SNP. This region is predicted to contain 10 genes (Figure 6c).

We identified three common haplotypes for this genomic region in the DartMap panel. Accessions that share the haplotype with the reference variant associated with SNP5^{7856132} will be referred to as haplotype 1, whereas accessions containing the non-reference variant can be further distinguished into two haplotypes that will be referred to as haplotypes 2 and 3 (Table S7). Haplotype 1 associates with higher ΦPSII in Fe-deficient conditions relative to the non-reference haplotypes, which do not significantly differ in ΦPSII (Figure 6d). The higher ΦPSII of accessions with haplotype 1 correlates with a significantly higher projected leaf area in Fe-deficient conditions than haplotypes 2 and 3 (Figure 6e), while both contain no significant differences in projected leaf area between each other.

In total, 40 sequence variants are in LD with SNP5^{7856132} at a cut-off of $r^2 > 0.55$, of which 27 are intergenic, 7 intronic and 6 exonic (Table S8). Only two of the exonic variants encode a non-synonymous amino acid substitution, but in both cases the substitutions are from a leucine to an isoleucine or vice versa. These are very similar amino acids, and neither of these substitutions, one in *RIBF1* (*At5g23330*), the other in *DUF789* (*At5g23380*), change the domains predicted by InterPro (Mitchell *et al*., 2019).

We thus hypothesized that non-coding variants that affect either levels or timing of gene expression should be causing the observed QTL. We measured expression of each candidate gene by means of RT-qPCR in both control and Fe-deficient conditions in shoots of six accessions of each haplotype (Table S8). Two known marker genes for Fe deficiency in leaves, *FSD1* and *FRO3* (Waters *et al*., 2012), indeed showed the expected lower (*FSD1*) or higher (*FRO3*) expression under the Fe deficiency treatment compared to control (Figure S8a). Two of the 10 candidate genes, *FSD3* and *STE14A* (*At5g23320*), are about 2-fold higher expressed in accessions with haplotype 1 compared to accessions with either of the other two haplotypes (Figure 6f and S8b). The ΦPSII in Fe deficiency correlates well with expression of these genes (Figure 6g). None of the eight other candidate genes showed such response of expression between haplotypes and correlation with the ΦPSII in Fe deficiency (Figure S8b).

Since the absolute expression levels of *STE14A* are very low, we focused on *FSD3* as the most likely candidate to underlie the QTL. Two distinct splice variants are described for this gene that influence functionality regarding chloroplast development (S. Lee *et al*., 2019). We measured the relative gene expression of each splice variant independently, but found highly similar splice variant ratios between the different haplotypes (Figure S8c), indicating that natural variation affects gene expression of *FSD3* only quantitatively and not through altering differential splicing. Analysis of the non-coding sequences surrounding *FSD3* identified a variable $(TTC)_n/(GAA)_n$ microsatellite repeat directly downstream of the *FSD3* 3′ UTR, which lacks either two or four of such repeats in haplotypes 2 and 3 relative to haplotype 1. As there are no differences in either ΦPSII or *FSD3* expression between haplotype 2 and 3 accessions we distinguished two alleles for the *FSD3* locus based on their functionality in Fe deficiency: an Fe deficiency tolerant allele, which has the Col-0 reference sequence, and a sensitive allele lacking two or four $(TTC)_n/(GAA)_n$ repeats. We subsequently confirmed the causal nature of the allelic variation at *FSD3* to explain the QTL using quantitative complementation, in which we test if the difference in ΦPSII between F_1 progeny derived from crosses between accessions with a tolerant allele crossed with both Col-0 and the *fsd3* mutant, and F_1 progeny derived from similar crosses with accessions carrying the sensitive allele. In Fe-deficient conditions, the F_1 progeny derived from the *fsd3* mutant and accessions with a tolerant *FSD3* allele significantly outperform the F_1 progeny derived from accessions with a sensitive *FSD3* allele ($P = 0.046$) (Figure 6h), but not under control conditions (Figure 6i).

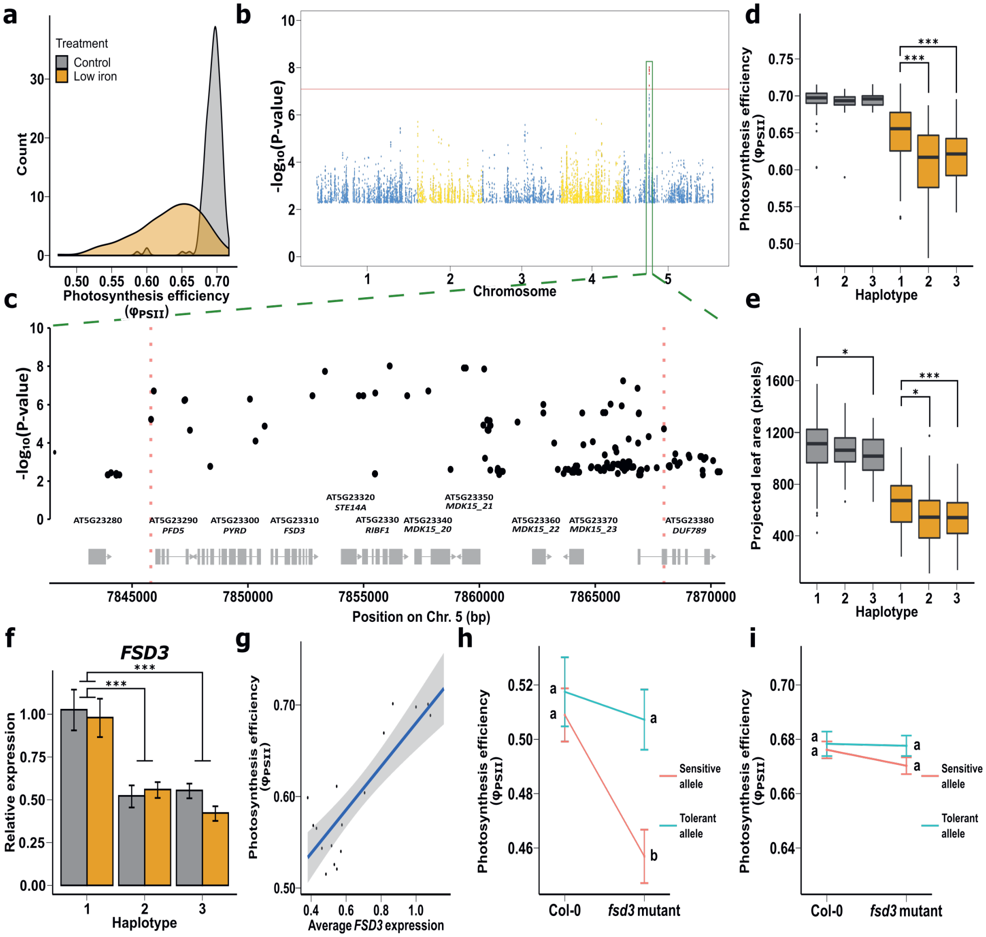


Figure 6: **Genome Wide Association (GWA) analysis of the response of photosynthesis efficiency to iron deficiency.** (a) Distribution of ΦPSII, the light-use efficiency of photosystem II (PSII) electron transport, when grown in control (21 μM Fe^{2+}, grey) and low iron supply (1 μM Fe^{2+}), orange). (b) Manhattan plot of the SNP variant associations [-log(P) > 2] when using the residual values of the linear regression of ΦPSII in control conditions versus iron deficiency conditions as trait for GWA analysis. The red line indicates a permutation-based threshold level of significance of P < 0.05. (c) The genomic region surrounding the significantly associated region on chromosome 5 (green box in (b)). Red dotted lines demarcate the area in linkage disequilibrium with the most significant SNP, containing candidate genes depicted below (in grey). (d-e) Phenotypic distributions for ΦPSII (d) and projected leaf area (e) per haplotype. (f) Relative gene expression levels per haplotype for *FSD3*. (g) Photosynthesis efficiency relative to the average relative *FSD3* expression level per accession with a linear fit (*FSD3*: R^2 = 0.74, P = 2.94 × 10^{-6}). (h-i) ΦPSII of F_1 progeny of crosses between Col-0 (tolerant *FSD3* allele) or the *fsd3* knockout mutant with accessions homozygous for the tolerant allele (N = 5) (blue) and similarly with accessions homozygous for the sensitive allele (N = 9) (red) in iron deficiency (h) and control conditions (i). Error bars represent SE. *P < 0.05; **P < 0.01; ***P < 0.001, two-tailed Student's t-test.

Discussion

Here we present the DartMap panel, consisting of 192 Dutch *A. thaliana* accessions and demonstrate its potential for detecting local adaptation. The genetic variation observed in this panel largely falls within the reported diversity of the 1001G collection, which is expected for accessions collected in Northwestern Europe outside the European relict regions (Alonso-Blanco *et al*., 2016). However, the genetic diversity is surprisingly high considering the relatively small geographical area without remarkable environmental clines. It illustrates the high level of standing genetic variation that exists among local populations of this predominantly self-pollinating species.

Despite, or perhaps due to, the absence of strong environmental clines, the DartMap panel is very suitable for conducting GWAS. The genetic architecture underlying natural adaptation is strongly dependent on the local environment (Brachi *et al*., 2013; Fournier-Level *et al*., 2013). Consequently, adaptive loci may exhibit genetic or allelic heterogeneity in regions with contrasting climates, which hampers their detection through GWAS. A milder diversity in environmental conditions can help to alleviate this limitation, as exemplified by the successful identification of the well-known *FRI* gene, which controls flowering time variation. This gene was not previously detected by GWAS in other collections such as the Iberian collection (Tabas-Madrid *et al*., 2018) or the large 1001G panel (Alonso-Blanco *et al*., 2016), primarily due to allelic heterogeneity. As such, the DartMap should be regarded as a valuable complement to collections sampled from regions with contrasting climates, and not a replacement. With the availability of several regional diversity panels, it will be easier to identify such adaptive loci when taking into account that each region may favour its own adaptive variation.

Our finding that genes involved in disease resistance are overrepresented among those impacted by CNV is consistent with previous work. A similar phenomenon was found for defense-response genes in a global set of 1,301 natural *A. thaliana* accessions (Göktay *et al*., 2021) and this same category of genes was found to be enriched in regions highly rearranged between seven diverse *A. thaliana* genomes (Jiao & Schneeberger, 2020). These observations suggest that CNVs may act as a genomic response to biotic stress. The next step would be to investigate whether such structural variants correlate to variation in the prevalence and abundance of known pathogens of *A. thaliana* (Bartoli *et al*., 2018).

The prevalence of plants with short inflorescences and the predominance of a near-isogenic genotype at various dispersed locations, strongly indicate an adaptive advantage for *ga5* mutants in the coastal regions of the Netherlands. Allelic heterogeneity and strong selection of one successful, near-isogenic genotype represented multiple times in the DartMap panel was the reason we could not identify the *GA5* locus through GWAS, but only based on prior knowledge on possible causal loci determining plant semidwarfism (Barboza *et al*., 2013). Signatures of selection were previously identified in *GA5* among *A. thaliana* accessions collected globally (Barboza *et al*., 2013), but the environmental cue driving such selection remained elusive. Later, Luo *et al*., (2015) identified several *ga5* loss-of-function plants in two Alpine populations of *A. thaliana* (Luo *et al*., 2015) and noted a decrease in plant height with increasing altitude although the adaptive advantage to high altitude environments specifically was not conclusively demonstrated (Luo *et al*., 2015). Our findings demonstrate that high-wind treatment significantly impacts the growth vigour of tall Dutch *A. thaliana* accessions, whereas it has no significant effect on the vigour of predominantly coastally located semidwarf plants.

Arguably, these are plants growing under optimal water and nutrient supply. Under natural circumstances, with additional growth-limiting factors and probably periods with much stronger winds, this reduction in vigour in the tall accessions is likely to be more detrimental, and also likely to affect their fitness (Younginger *et al*., 2017), which we could not establish under our artificial windy conditions. The observed phenotypic responses in terms of biomass and stem length to windy conditions are similar between the natural Dutch accessions and the Col-0 and Ler reference strains, as well as their corresponding *ga5* loss-of-function lines. Considering that wind speed generally increases with altitude in mountainous areas (https://map.neweuropeanwindatlas.eu), the selective advantage of the Alpine semidwarfs is also likely to be due to tolerance to wind speed. The selective pressure seems high, given the many independent *ga5* alleles found in the Netherlands, and elsewhere in the world (Barboza *et al*., 2013; Luo *et al*., 2015). It is remarkable that the near-isogenic genotype collected at several Dutch sites spread over such large distances, without any obvious means of transport. This is in contrast to a previous case of whole-genome hitchhiking, due to selection on herbicide tolerance, where the spread was clearly linked to transportation along railway tracks (Flood, Van Heerwaarden, *et al*., 2016).

The presence of a North-South cline in flowering time and the association between windy conditions and inflorescence height both highlight the importance of relatively mild environmental clines for local adaptation. These patterns can be detected in a uniformly and densely sampled population like the DartMap. This is further exemplified by the strong association on chromosome 4 related to the mild cline in seasonality, where the alternative allele of *qDPT-B* is only found at a high local abundance in the West of the Netherlands. However, it should be noted that such associations do not necessarily provide evidence of causality. Demographic processes or genetic drift can produce similar patterns to those created by selection (Rellstab *et al*., 2015). To further investigate whether this association represents a case of local adaptation, common garden experiments or reciprocal transplanting experiments can be conducted (de Villemereuil *et al*., 2016; Hancock *et al*., 2011; Rellstab *et al*., 2015; Terés *et al*., 2019). Nonetheless, the high local relative abundance observed in the West of the Netherlands compared to rest of the world, along with the differential sensitivity to drought, suggests the presence of an adaptive locus. However, it seems unlikely that this adaptation is solely attributed to the drought response measured in our experiments, as similar climate characteristics in other regions would also be expected to exhibit a higher relative abundance. Therefore, it is more plausible that a combination of different climatic factors and/or other unconsidered factors such as urbanization contribute to this adaptation.

Identifying the causal gene in genotype-environment association analysis can be challenging because allelic variation may affect tolerance to an experimentally measured trait that may not necessarily reflect the trait under natural selection. However, in our study we found that the mutant of *La-RELATED PROTEIN 1c* (*LARP1c*) is more sensitive to drought compared to other tested mutants and Col-0. *LARP1c* encodes a protein with a conserved La-motif domain, which is associated with RNA recognition and is widely found among eukaryotes (Bayfield *et al*., 2010). In *A. thaliana*, *LARP1c* is known to be involved in seed germination and the regulation of leaf senescence, which can be induced by various environmental stresses (Yan *et al*., 2023; B. Zhang *et al*., 2012). Additionally, each of the three LARP1 (1a, 1b, 1c) proteins in *A. thaliana* localise at the processing bodies (Yan *et al*., 2023; B. Zhang *et al*., 2012) and are proposed to play a role in RNA processes in response to environmental perturbations. For instance, LARP1a is required for thermotolerance of *A. thaliana* to long exposure of moderately high temperatures, where it is involved in regulating the RNA decay machinery together with the cytoplasmic exoribonuclease *XRN4* (Merret *et al*., 2013). However, the specific functions of *LARP1c* in plants are still relatively understudied, and there may be other roles yet to be explored (Yan *et al*., 2023). *LARP1c* may thus also be involved in the response to drought and possibly also other environmental factors.

In addition to the mild and gradual environmental clines, we also examined the panel for sensitivity to low iron availability, as we expected to find variation for this trait given the distinct differences in soil type across the Netherlands. This identified non-coding variation at *FSD3* to cause a major QTL affecting sensitivity to low iron supply. *FSD3* is long known to encode an essential chloroplast-localized iron superoxide dismutase. Its function is to protect developing chloroplasts against excessive damage by superoxide (Myouga *et al*., 2008). However, natural genetic variation in *FSD3* has not been previously observed. The QTL affecting sensitivity to low iron supply is influenced by subtle allelic variation in *FSD3*, specifically related to the number of repeats in a microsatellite region directly downstream of the transcribed region. This variation results in a two-fold difference in gene expression between contrasting alleles, which strongly correlates with photosynthesis efficiency under iron-deficient conditions. When plants are supplied with sufficient iron, there is no phenotypic difference between contrasting haplotypes. *FSD3* is one of two chloroplast-localized iron superoxide dismutases, the other being *FSD2* (Myouga *et al*., 2008). While both can form heterocomplexes and *FSD3* overexpression can partly complement the *fsd2* mutant phenotype (Gallie & Chen, 2019), they are not genetically redundant. The *fsd2* loss-of-function mutant displays a compact rosette phenotype, with pale green leaves, and the *fsd3* loss-of-function mutant is barely viable, with a very small and pale rosette (Myouga *et al*., 2008). Since iron is an essential co-factor for iron superoxide dismutases, a decrease in iron supply is likely to affect their function. Under the mild iron deficiency conditions we tested, this decrease in function was apparently just enough to affect photosynthesis efficiency significantly more in the genotypes with lower *FSD3* expression.

We found no reason to suggest strong selection for either of the two alleles. Since *FSD3* function is essential for proper chloroplast development and tolerance to reactive oxygen species generated by photosynthesis, lower *FSD3* expression is likely disadvantageous. The impact of lower *FSD3* expression may even extent beyond chloroplast development and photosynthesis efficiency. For instance, hydrogen peroxide, which is produced when superoxides are converted by superoxide dismutases, can function as a signaling molecule in the systemic response to iron deficiency (Le *et al*., 2016). Therefore, plants with lower *FSD3* expression may have a delayed response in detecting iron deficiency. Nevertheless, a large ionome survey of 1135 global *A. thaliana* accessions did not reveal any clear indications of local adaptation to iron deficiency (Campos *et al*., 2021). This suggests that establishing a direct link between allelic variation and local adaptation to iron deficiency may be challenging. Factors such seasonal variation in iron availability depending on precipitation or limited distribution and heterogeneity of the allelic variation could contribute to the difficulty in establishing such a link.

In conclusion, our work demonstrates that regions with relatively mild environmental clines can still harbor a wide range of adaptive genetic variants. This highlights the value of establishing regional collections as excellent tools for studying local adaptation of *A. thaliana*. Such a collection does not need to be limited to regions with strong environmental clines as were put forward in previous studies (Brachi *et al*., 2013; Frachon *et al*., 2018). Even a diversity panel from a relatively small geographic region with mild environmental clines, such as the Netherlands, provides enough variation to pick up small differences in flowering time and sensitivity to iron deficiency. Given the decreasing costs of whole genome sequencing, it becomes increasingly easier to characterize the genetic variation in such populations in great detail. Our study may serve as a blueprint of this approach, supporting the view that collections of wild plant species can be powerful systems to determine the genomic basis of natural adaptation (Monnahan *et al*., 2019; Sollars *et al*., 2017; Weigel & Nordborg, 2015).

Materials and Methods

Plant material

A set of 192 lines was selected from a large collection of around 2,000 natural accessions sampled throughout the Netherlands, with the aim to obtain a uniform geographical spread of accessions and generate a local diversity panel, the DartMap panel, suitable for GWAS (Supplementary Table 9). While accessions were selected from a larger set of about 2,000 collected individuals, the selected panel has a relatively lower density of accessions in the north of the Netherlands and in the very south, due to limited availability or collector capacity. Seeds collected from individual plants in the field were used to propagate each accession as a line for two generations through single-seed-descent upon self-fertilization in a greenhouse, before DNA isolation.

DNA isolation

Genomic DNA was extracted from one or more inflorescences (open flowers and above) from a single plant per accession. The material was frozen in liquid nitrogen, ground to a fine dust and incubated in 300 µL 2x CTAB buffer (2% CTAB, 1.4m NaCl, 100mm Tris, 20mm EDTA, pH 8) for 30 minutes at 65 °C. An equal volume of chloroform was added, mixed, centrifuged (3250 rpm for 15 min) and the supernatant was collected. The DNA was precipitated by adding an equal volume of ice-cold isopropanol, incubated overnight at -20 °C and centrifuged (3250 rpm for 15 min). The precipitate was washed twice with 70% ethanol and air dried. DNA was dissolved in milliQ water and treated with RNase (Promega) for 30 minutes at 37 °C.

Genomic DNA sequence data

Genomic DNA of the DartMap panel was sequenced at GenomeScan B.V., Leiden. Libraries were prepared using the NEBNext Ultra DNA Library Prep kit for Illumina, according to described procedures. 500-700 bp insert size fragments were sequenced using a Illumina Hiseq 4000 device, aiming for at least 30x coverage of paired-end reads (2x151 bp) of each accession, and used to generate the alignment files necessary for calling genomic variants (see Supplementary Methods for details). We additionally generated alignments for the DartMap panel combined with 11 Dutch accessions included in the 1,001 Genomes (1001G) Project (Alonso-Blanco *et al*., 2016) (Sequence Read Archive ID: SRP056687). This combined set of samples will be referred to as the "DartMap + 1001G panel", to clearly distinguish it from the DartMap panel containing accessions collected in this study only.

Single nucleotide polymorphism (SNP) and indel calling

Genetic variants were called using a workflow based on the Genome Analysis Toolkit (GATK) Best Practices (Van der Auwera *et al*., 2013). Base quality scores of aligned reads were recalibrated using GATK (version 4.0.2.1) BaseRecalibrator with default parameters, using a set of variants called in a world-wide panel of 1135 *A. thaliana* accessions as known sites (Alonso-Blanco *et al*., 2016) (obtained from ftp.ensemblgenomes.org/pub/plants/release-37/vcf/arabidopsis_ thaliana). SNPs and short insertions/deletions (indels) were called in each sample using GATK HaplotypeCaller, allowing a maximum of three alternate alleles at each site. Samples were jointly genotyped using GATK GenomicsDBImport and GATK GenotypeGVCFs with default parameters. This last step generated three different VCF files containing calls for the nuclear genome, the mitochondrial genome, and the chloroplast genome respectively. We filtered these sets to remove putative false positive calls (see Supplementary Methods for details).

Copy number variation (CNV) calling

We called CNVs in the DartMap panel, defined here as deletions, duplications, and insertions of at least 50 bp, by applying the Hecaton workflow (Wijfjes *et al*., 2019) to the alignment files of each accession. Accessions of the 1001G were excluded from this analysis, as they were sequenced using a different sequencing platform, which can strongly affect CNV calling: a linear model of genetic distance between accessions based on CNVs as a function of differences in coverage, sequencing technology, insert size, and read length (see Supplementary Methods for details) explained 30% of the variation in the number of CNVs found in each sample (R^2 = 0.3). CNVs called in each accession were genotyped and merged to generate a single VCF file as output (see Supplementary Methods for details).

Characterizing the impact of deletions

We detected deletions overlapping protein-coding genes (by at least 1 bp) using bedtools intersect. Genes were obtained from the TAIR10 genomic annotation of Ensembl Plants (release 40) and from the annotation of the improved *A. thaliana* mitochondrial assembly (BK010421.1). Gene function was considered disrupted if the associated gene overlapped with a deletion and at least 76 bp (more than half the length of a read) of the coding sequence lacked read coverage. Deletions were defined as having a moderate frequency if they were present in at least 48 samples (25%) of the DartMap panel. We defined disease resistance genes by first collecting all Uniprot entries linked to all *A. thaliana* reference genes using the Retrieve/ID mapping tool (https://www.uniprot.org/uploadlists/), excluding a small number of genes (53) for which no corresponding entries could be found. Genes were labeled as involved in disease resistance if the name of their protein product contained the terms “disease resistance” or “Disease resistance”. Enrichment tests were performed using Fisher’s exact test as implemented in the Python SciPy library (Virtanen *et al*., 2020) (version 1.3.1).

Population structure analysis

We clustered samples based on bi-allelic sites missing calls in fewer than 10% of the samples and having a non-reference allele in at least 10 of them, to ensure that accessions are grouped on variation at the population level, rather than rare variants found in few individuals. For clusters based on mitochondrial and chloroplast variants, this criterium was relaxed to a minimum of two samples, as a relatively large number of sites would have been discarded otherwise. To generate a distance metric, we computed the Hamming distance between samples represented by binary vectors with 1s for homozygous variants and 0 for all other cases, using the dist. gene function of the R package ape (v5.3) (Paradis & Schliep, 2019). Samples were clustered with the hclust function in R (R Core Team, 2013), using complete linkage as the agglomeration method, similar to analyses on the Swedish and Iberian collection (Long *et al*., 2013; Tabas-Madrid *et al*., 2018) (Figure S9). We cut trees into different numbers of groups using the R function cutree, to explore the geographical distribution of clusters formed at various hierarchical levels. Geographical distance between accessions was computed using the distance function of the geopy library (https://github.com/geopy/geopy) (v2.0.0) with default parameters and their geographical locations were plotted using the basemap library (https://matplotlib.org/basemap/) (v1.2.2).

Principal component analysis

Principal component analysis of chloroplast variants called in the Dutch and global collection was performed using the SeqArray (v1.28.0) (Zheng *et al*., 2017) and SNPrelate (v1.22.0) (Zheng *et al*., 2012) libraries in R. To remove putative false positive calls, variants were excluded if they had a quality-by-depth (QD) score lower than 25, leaving a total of 4095 sites. Furthermore, we removed 10 accessions in which at least 50% of the variants were called as heterozygous, suggesting DNA contamination of the samples used for sequencing. Results were visualized using ggplot2 (v3.3.2) (Wickham, 2016). Three outlier accessions were excluded from the final figures to improve the visualization of Dutch accessions with respect to the global collection. Accessions were coloured based on their chloroplast haplogroup assigned in a previous study (Hsu *et al*., 2019). Haplogroups of accessions that were not included in that paper (i.e. the DartMap panel) were predicted with a 1-nearest neighbour classifier on the accessions that were, using the first two principal components as features.

Phenotyping flowering time and stem length

For all experiments, seeds were sown on wet filter paper and stratified for 3 days at 4 °C before they were germinated. Seedlings were grown hydroponically on rockwool blocks (Grodan Rockwool Group, 40 x 40 mm in size) pre-soaked in a nutrient solution designed for *A. thaliana* (1.7 mM NH_4^+, 4.5 mM K^+, 0.4 mM Na^+, 2.3 mM Ca^{2+}, 1.5 mM Mg^{2+}, 4.4 mM NO_3^-, 0.2 mM Cl^-, 3.5 mM SO_4^{2-}, 0.6 mM HCO_3^-, 1.12 mM PO_4^{3-}, 0.23 mM SiO_3^{2-}, 21 µM Fe^{2+} (chelated with 3% diethylene triaminopentaacetic acid), 3.4 µM Mn^{2+}, 4.7 µM Zn^{2+}, 14 µM BO_3^{3-} and 6.9 µM Cu^{2+} at pH 7, EC 1.4 mS cm^{-1}). Plants were grown in a greenhouse until 97 days after sowing. For each treatment, plants were equally divided over 20 plastic trays, with 76-78 plants per tray. Eight replicates per genotype per treatment were used, randomly assigned over four separate trays with two replicates per tray per genotype. Each tray consisted of a unique, randomly generated combination of genotypes. Trays were placed in a random order in the greenhouse, and were moved once every two weeks to a new position to minimize possible spatial effects in the greenhouse. A vernalisation treatment of six weeks was performed in a climate-controlled growth chamber set at 4 °C, with 12h/12h light/dark period, 70 $\mu m \cdot m^{-2} \cdot s^{-1}$ irradiance and 70% humidity starting at day 20 after sowing. After vernalisation, plants were returned to the greenhouse. Flowering time and height up to the first silique were scored in the greenhouse, either with or without a vernalisation period of six weeks. Flowering time was determined as the day when the first flower had opened. Stem length up to the first silique was measured one week after the first flower had opened, as this trait remains stable thereafter (Barboza *et al*., 2013). To obtain the set of phenotypes for further analysis, we calculated the average phenotypic values per accession per treatment.

Wind simulation treatment

In a separate experiment, plants of six semidwarf accessions and five tall accessions of the DartMap panel, with similar flowering time, as well as the accessions Col-0 and Ler, and their respective loss-of-function *ga5* mutants (Supplementary Table 3) were grown for the first 14 days after sowing in a climate-controlled growth chamber, set at 20/18 °C (day/night) with 10 h light (from 08:00 - 18:00) and 70% humidity at an irradiance of 200 µm · m^{-2} · s^{-1} at ambient CO_2 levels. The growth chamber contained two treatment basins. The treatment basins have the same design as described in a previous study (Flood, Kruijer, *et al*., 2016), with the distinction that the two treatment basins are physically separated and are on either side of the climate chamber. Each treatment basin was divided into five equally sized blocks to hold plants. Accessions were grown in a randomized complete block design with six replicates per block. Wind was simulated with two fans that were positioned at either end of one of the treatment basins. The fans were set to provide a wind speed that varied around 5.6 m/s for the majority of the treatment basin. The wind speed varied depending on the distance of the plant to the fan, with plants closest to the fan experiencing wind speeds up to 7 m/s and plants furthest away of the fan and at the edge of the treatment basin experiencing 3.5 m/s wind. The wind speed for the control treatment basin varied between 0.5 and 1.0 m/s. Wind speeds were measured with a TSI Velocalc (model: 8347-M-GB) air velocity monitor. Wind treatment was simulated by alternating patterns of six hours with wind by six hours without wind and was terminated when the first plants in the wind treatment basin had fully ripened siliques.

Extraction of climatic variables

Climatic variables were obtained from google earth engine using R and RGEE (Aybar *et al*., 2020). All bands from Worldclim BIO data (Hijmans *et al*., 2005) selected from 250m-by-250m squares using the coordinates from the collection site for each accession as input.

Drought treatment

To test which of the two nearby QTLs at *qDPT* affected drought tolerance, plants were grown in a greenhouse on ceramic-based granular soil. For the first two weeks, all plants were watered three times a week with the nutrient solution described previously. This watering regime was continued for the control treatment while for the drought treatment excess nutrient solution was removed and watering was ceased. Accessions (Table S5) were grown in a randomised complete block design with five replicates per treatment. Photographs were taken with a Nikon D3000 with the same settings, zoom and distance to the plants and were subsequently analysed to obtain the projected leaf area (PLA) with ImageJ (settings Hue 26-114, Saturation 58-218 and Brightness 121-255). The PLA was then normalised per accession by dividing the PLA in the drought treatment by the PLA in the control treatment to correct for innate differences in size of different accessions. For candidate gene validation T-DNA insertion lines were used

(Table S6). Plants were grown in our phenotyping growth chamber, the Phenovator (Flood, Kruijer, *et al*., 2016), with the same conditions in the growth chamber as described for the wind simulation treatment. The drought treatment was induced in a similar way, by using ceramic-based granular soil. In contrast to the experiment in the greenhouse, the growth substrate was covered with black plastic cover to prevent algal growth that might interfere with high throughput phenotyping. Moreover, we watered plants in the drought treatment only once at the start, as there was less evaporation in this setup compared to the greenhouse. Starting at 10 days after sowing, we measured projected leaf area based on near infrared reflectance (NIR) imaging as described by Flood, Kruijer, *et al*., 2016 *et al*., (2016) four times per day, and Fv/Fm using chlorophyll fluorescence once per day.

Iron deficiency treatment

In a separate experiment, plants were grown as described for the wind simulation experiment, but in this case, seven rather than five blocks per treatment basin were used. Plants in the control treatment were watered with the nutrient solution described above, while those under iron deficiency treatment received an adjusted solution with only 1 µM Fe^{2+} instead of 21 µM Fe^{2+}. Either three or four replicates were used per accession per treatment, with half of the accessions having four replicates in control condition and three replicates in the iron deficiency treatment and vice versa for the other half of the accessions. On day 15, plants were transferred to our phenotyping growth chamber, the Phenovator (Flood, Kruijer, *et al*., 2016), with the same conditions. We measured the steady state quantum yield of photosystem II electron transport (ΦPSII), as a proxy for photosynthesis efficiency, and the projected leaf area, based on near infrared reflectance (NIR) imaging as described by Flood, Kruijer *et al*., (2016), at five resp. seven time points per day. We discarded outlier samples with NIR values smaller than the mean minus two standard deviations per treatment. We corrected the raw phenotypic data (ΦPSII and projected leaf area) for block effects by fitting the following linear mixed model in which random terms are underlined:

$$\underline{Y} = \text{Genotype} + \text{Treatment} + (\text{Genotype} \times \text{Treatment}) + \underline{\text{Block}} + \underline{\varepsilon}$$

The model was analysed using the restricted maximum likelihood procedure with the lme4 package in R (Bates *et al*., 2014). For the GWAS, we calculated the average ΦPSII per accession per treatment. To obtain a single phenotypic value representing a measure of tolerance per accession, we then plotted ΦPSII for each accession under the iron deficient condition versus control condition and fitted a linear regression through the entire population. We calculated the residuals of each accession relative to this linear regression to be used as the phenotypic value for GWA analysis. Additionally, we used the average ΦPSII values in the iron deficient condition and the calculated ratio between ΦPSII in iron deficiency versus control as alternative traits for GWAS.

Plant material for expression analysis

Plants that were used for gene expression analysis were grown in the greenhouse on rockwool blocks that were watered with either the control nutrient solution (21 µM Fe^{2+}) or the adjusted low iron solution (1 µM Fe^{2+}). For both treatments we used five plastic trays to contain the nutrient solution. Two replicate plants per accession were distributed over trays in a complete randomised block design. Whole shoots of two replicate plants per accession originating from the same tray were pooled and harvested as one sample for RNA isolation 21 days after germination, while five such samples per accession per treatment were harvested in total. The material was snapfrozen in liquid nitrogen immediately after harvesting and stored at -80 °C until RNA isolation. Accessions were selected for gene expression analysis (Supplementary Table 7) based on their haplotype for the region with the highest association in the GWAS.

Quantitative complementation

Quantitative complementation was used to test if the difference in ΦPSII between the F_1's of Columbia-0 (Col-0) and the knockout mutant *fsd3* in Col-0 background crossed with accessions homozygous for the tolerant allele was significantly different from the response in ΦPSII for F_1's of Col-0 and *fsd3* crossed with accessions homozygous for the sensitive allele. Five accessions with a tolerant, and nine accessions with a sensitive allele for *FSD3* were crossed to both Col-0 and *fsd3* (Table S7). Col-0 and *fsd3* were used as pollen donor in all crosses. The *fsd3* T-DNA insertion knockout mutant, SALK_103228, was ordered from the Arabidopsis Biological Resource Center and insertion of the T-DNA was validated by PCR as earlier described (Myouga *et al*., 2008). This T-DNA insertion line has no *FSD3* transcript (Myouga *et al*., 2008), and we could only grow it beyond the seedling stage on a 0.3% gel-rite medium supplemented with 3% sucrose.

Plants (N = 10 per cross per treatment) were grown as described in the section "Plant material for expression analysis". Plants were phenotyped for ΦPSII from 21 DAS onwards as described by Baker and Oxborough, 2004 with the PlantScreenTM SC System (Photosystems Instruments). Phenotyping was done for twenty plants per run and lasted for three consecutive days overall. We phenotyped crosses with the tolerant allele and the sensitive allele equally spread out over the three days to prevent possible phenotypic differences caused by the delay between measurements. Moreover, in each measurement of twenty plants we included F_1's from both the cross between a given accession with Col-0 and with *fsd3* in a single measurement. We included one reference tray that was measured each day in the morning, at the start of the afternoon and at the end of the afternoon to test for possible differences in ΦPSII during the day, but no significant differences were found. We corrected the phenotypic data for possible block effects by fitting a linear model, with random terms underlined:

$$\underline{Y} = (\textit{FSD3}\text{ allele group x Pollen donor}) + \underline{\text{Block }} + \underline{\varepsilon}$$

The model was analysed using the restricted maximum likelihood procedure with the lme4 package in R (Bates *et al*., 2014). We tested interaction between the *FSD3* allele group (natural tolerant or sensitive allele) and the pollen donor (Col-0 or the *fsd3* mutant) with a two-way ANOVA, for which a Kenward-Roger approximation for the degrees of freedom was used. Tukey's post-hoc multiple comparison test (α = 0.05) was used to assess significance.

RNA isolation and gene expression analysis

Total RNA was extracted with the Direct-zol RNA mini-prep kit (Zymo Research) according to the company's instructions and then subjected to a DNAse (Promega) treatment at 37 °C for one hour. Total removal of DNA was validated by means of a no-reverse-transcriptase PCR reaction on a few mixtures of samples. The RNA quality was assessed for purity (A260/A280) with a spectrophotometer (Thermo scientific Nanodrop 2000) and for possible RNA degradation by means of a visual inspection of the RNA on a 1% agarose gel. cDNA was then synthesized from 1 µg total RNA (measured by spectrophotometer) with the iScript™ cDNA synthesis kit (Bio-RAD) according to the company's instructions. Reverse transcriptase quantitative PCR (RT-qPCR) reactions were performed with the SensiFast SYBR Green Mastermix (Bioline) on a Bio-RAD cfx96 machine. For each RT-qPCR primer set used, we first validated its efficiency by means of a standard curve and only used primer sets with efficiencies ranging between 90% and 110%. Gene expression was normalized, using the delta cycling threshold (ΔCt) method (Livak & Schmittgen, 2001), to three reference genes, *SAND, YLS8*, and *TIP41-like* as earlier described by Han *et al*., 2013. Differential expression was assessed by two-sample t-tests on the ΔCt values. All primer sequences can be found in Supplementary Table 10. We included *FSD1* and *FRO3* in the expression analysis as marker genes for iron deficiency in shoots (Waters *et al*., 2012).

Genome-wide association (GWA) and genome-environment association (GEA) analyses

GWA and GEA mapping were performed in GEMMA (Zhou & Stephens, 2012) (version 0.98) with a univariate linear mixed model and the minor allele frequency cut-off set at 0.05. A kinship matrix was constructed in GEMMA to correct for population structure. CNVs were included in the analysis in a similar fashion as SNPs and indels by converting them into a reference variant, an alternative variant or heterozygous. To estimate a suitable genome-wide *p*-value threshold, we performed 1,000 permutations of the phenotype values. For each permutation we performed association analysis and extracted the highest association score. The distribution of these 1,000 highest association scores was then used to determine the empirical threshold value of $p < 0.05$.

Data availability

Raw sequencing data of the DartMap panel can be found on the repository of the National Center for Biotechnology Information (NCBI) (BioProject ID: PRJNA727738).

Acknowledgements

We thank Yanda Zhou, Jeroen van Buren, and Imme Bartels for their contributions to the data collection of phenotyping experiments; the Unifarm staff for taking excellent care of all plants grown during this work; and Korbinian Schneeberger for providing feedback on the manuscript. We are especially grateful to Erik Wijnker and Joost Keurentjes, who called upon the listeners of the Dutch radio program "Vroege Vogels" (https://www.bnnvara.nl/vroegevogels) to collect *A. thaliana* plants and seeds and send them to Wageningen University, and all those people who did. Without them, we would not have been able to get such a well-distributed set of Dutch *A. thaliana* accessions.

Supplementary Figures

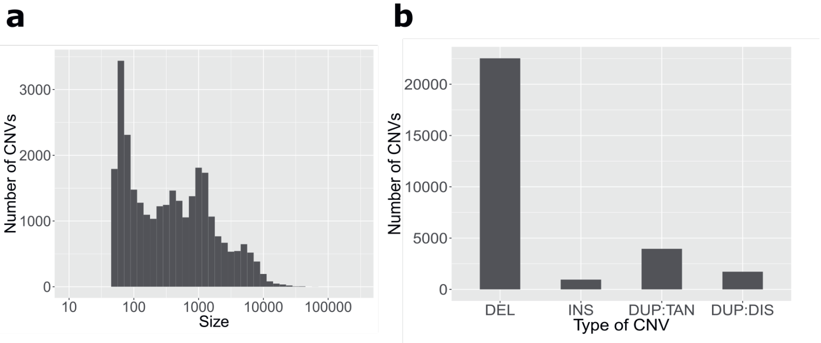


Figure S1: **CNVs in the DartMap panel.** (a) Size distribution of CNVs. CNVs below 50 bp are excluded, as these are considered indels. (b) Type distribution of CNVs: deletions (DEL), insertions (INS), tandem duplications (DUP:TAN), or dispersed duplications (DUP:DIS).

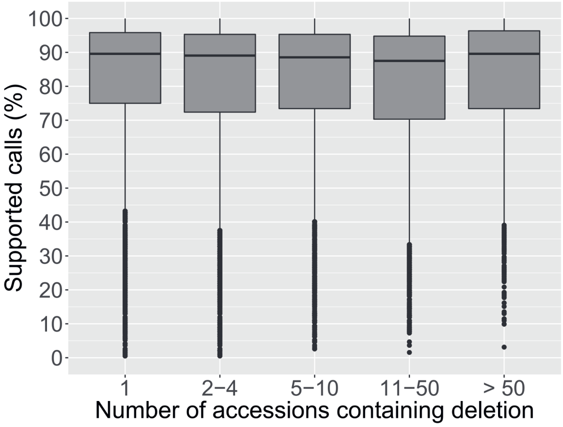


Figure S2: **Deletions in the DartMap panel were called with high precision.** The y-axis shows the percentage of genotype calls supported by a matching decrease (non-reference allele) or no change (reference allele) in read depth. Deletion variant sites are stratified along the x-axis by frequency of the non-reference (deletion) allele.

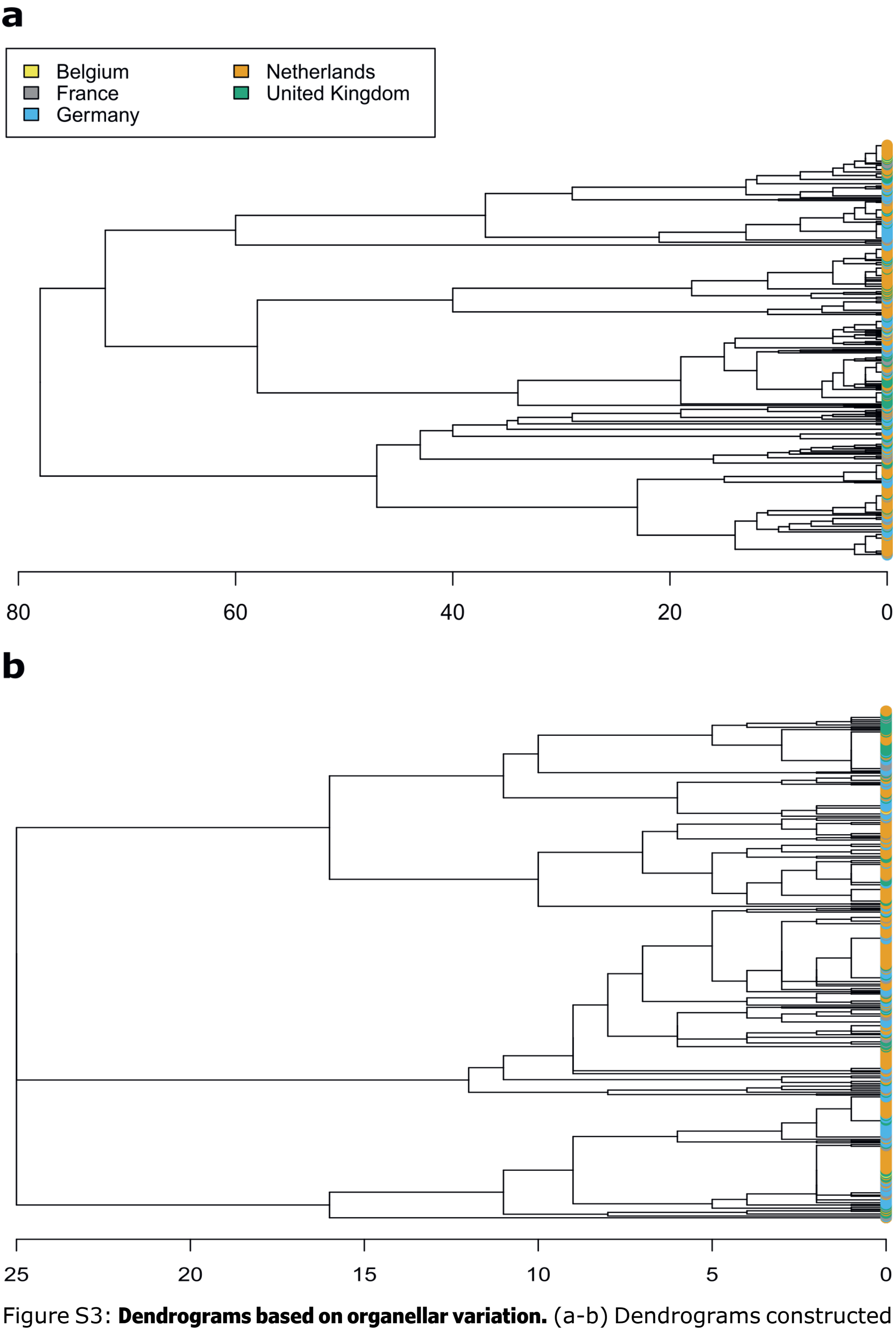


Figure S3: **Dendrograms based on organellar variation.** (a-b) Dendrograms constructed based on pairwise similarity of chloroplast (a) and mitochondrial (b) variants between Dutch accessions and those of nearby countries.

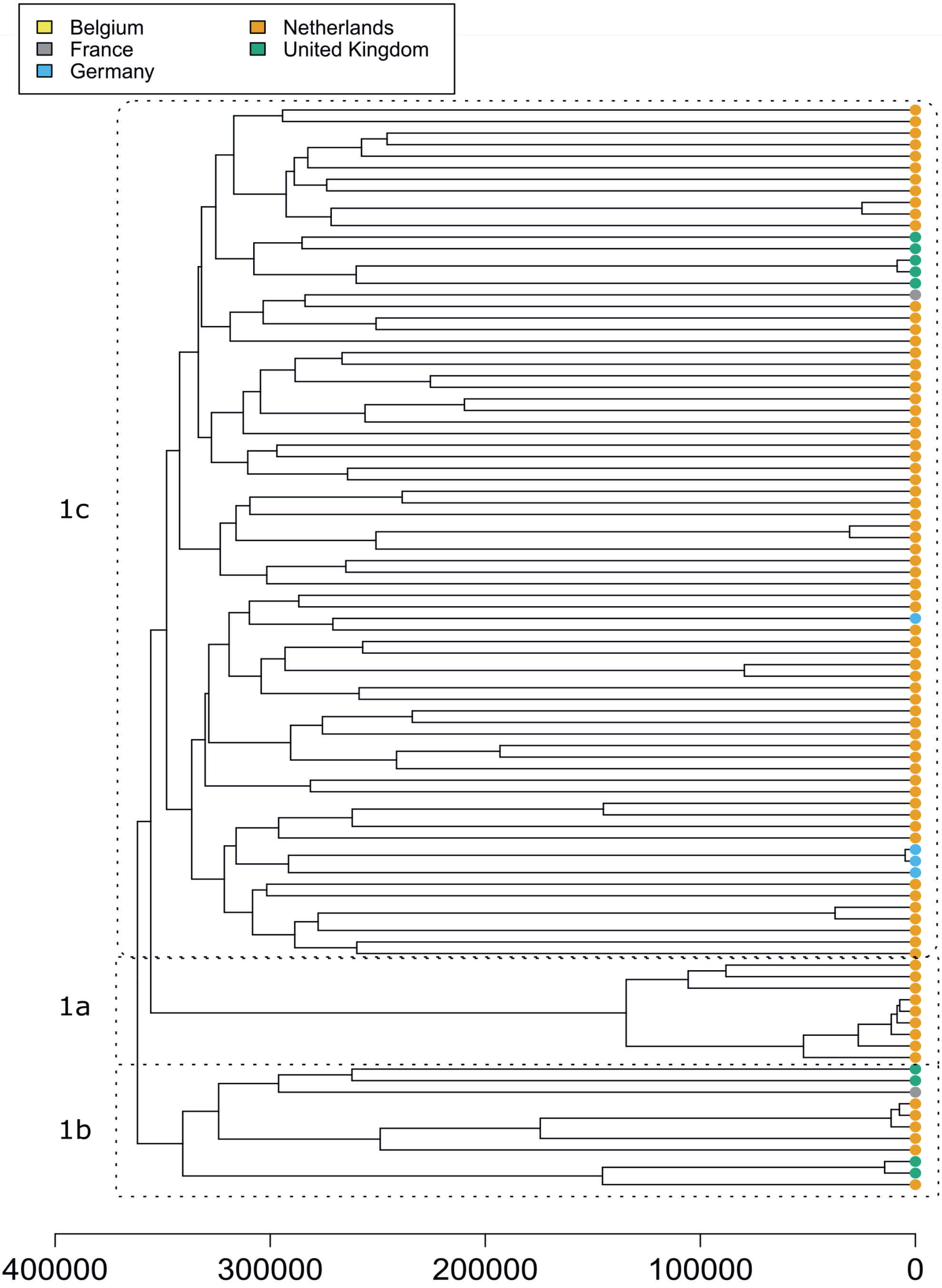


Figure S4: **Dendrogram of group 1 based on nuclear variation.** At its highest level, the dendrogram constructed based on pairwise similarity of nuclear variants between Dutch accessions and those of nearby countries can be split into two groups. Group 1 and its three subgroups (1a, 1b, and 1c) are depicted here.

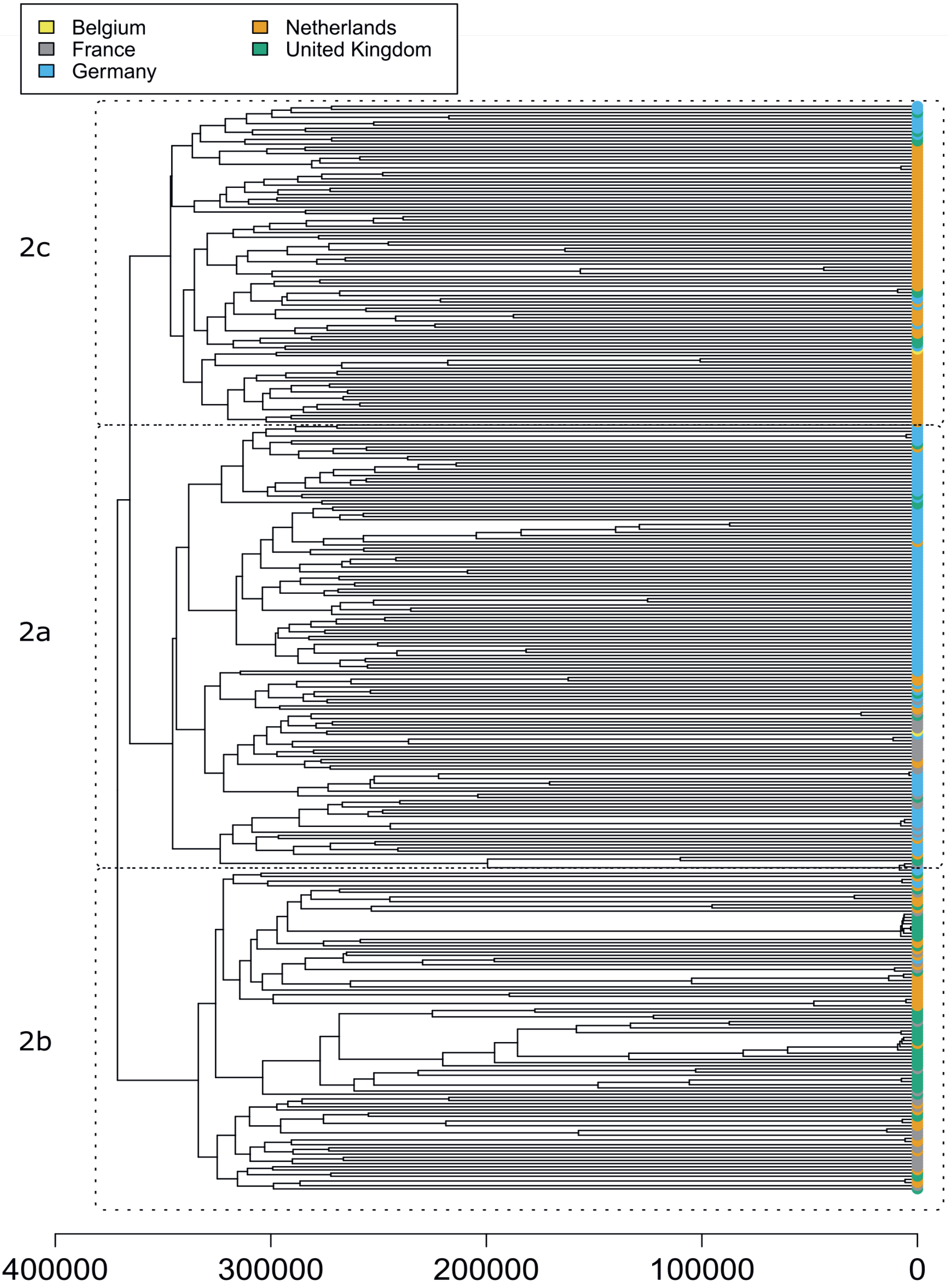


Figure S5: **Dendrogram of group 2 based on nuclear variation.** At its highest level, the dendrogram constructed based on pairwise similarity of nuclear variants between Dutch accessions and those of nearby countries can be split into two groups. Group 2 and its three subgroups (2a, 2b, and 2c) are depicted here.

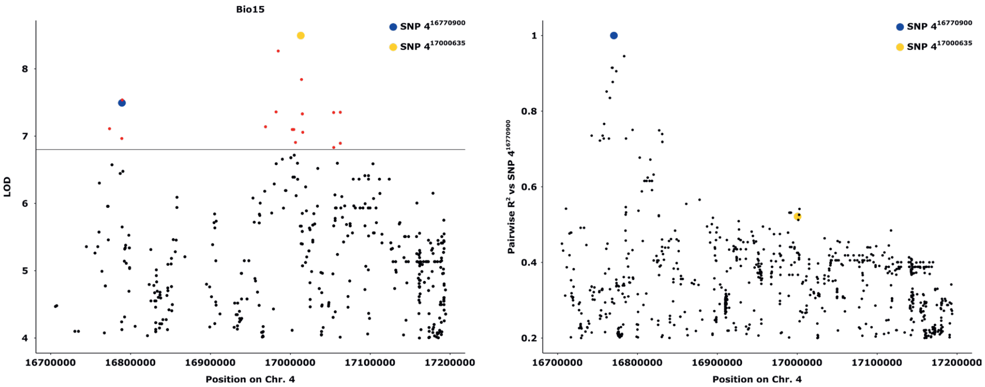


Figure S6: **Genomic region surrounding the chromosome 4 peak that was mapped for multiple climatic variables.** (a) Zoomed in section of the Chr. 4 peak of the mapping of precipitation seasonality (figure 5b) shows that there appear to be two nearby QTLs. The two SNPs that were found most frequently with the highest LOD score among the different mappings are depicted with a blue and yellow dot. b) Linkage analysis (expressed as the pairwise R^2) of SNPs in the Chr. 4 peak area relative to SNP $4^{16770900}$ (blue dot) reveals that the two QTLs are not in strong LD.

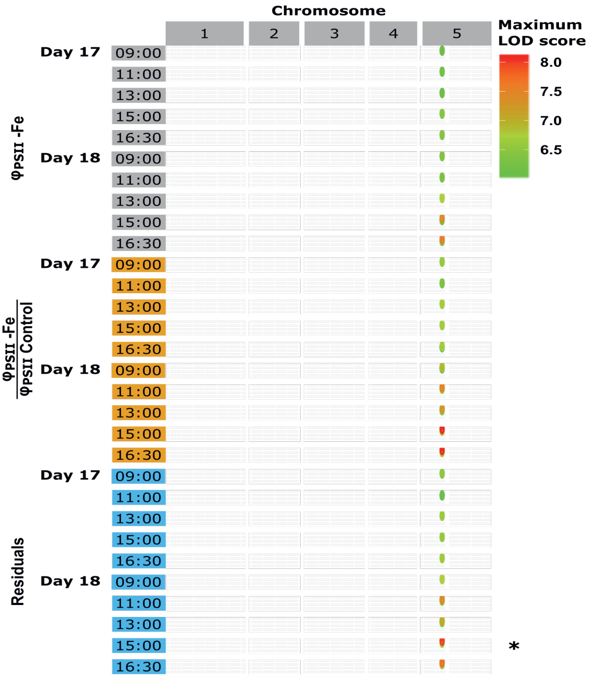


Figure S7: **Iron deficiency GWAS time series.** Summary data of the GWA results over 10 time points, measured during days 17 and 18 after sowing for three different phenotypes: [1] the average ΦPSII per accession in iron deficient conditions (grey), [2] the ratio between ΦPSII in iron deficient conditions and ΦPSII in control conditions (orange), and [3] the residuals (blue). The dots on the graph represent the highest LOD scores in a 25 kb window size above a LOD threshold of 6. The * represents the data from the Manhattan plot in Figure 6b.

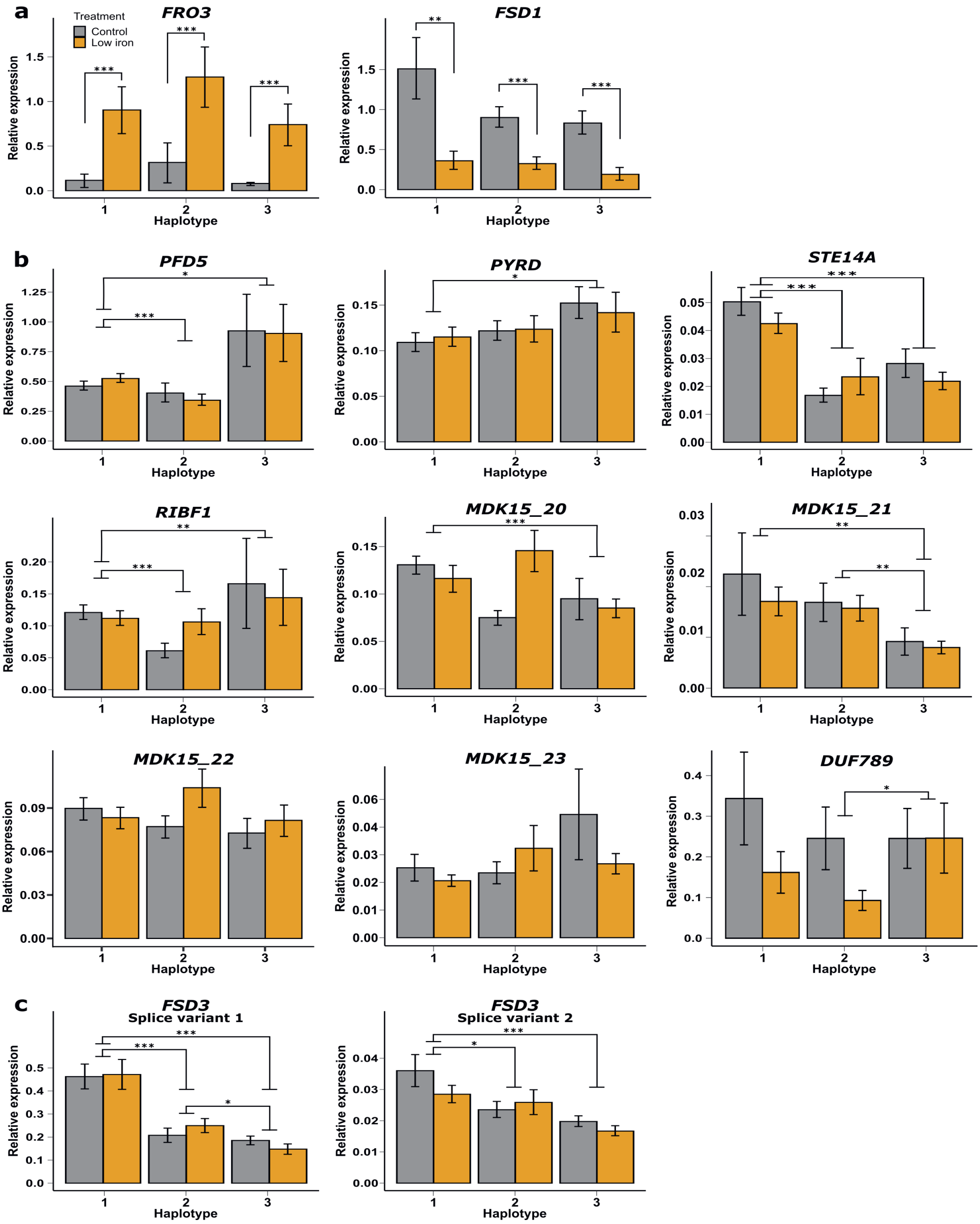


Figure S8: **Gene expression levels per haplotype group as measured by RT-qPCR.** The average gene expression levels at (21 µM Fe^{2+} (Control, depicted in dark grey) and (1 µM Fe^{2+}) (Low iron, depicted in orange) per haplotype group for two iron deficiency marker genes (a), nine candidate genes from the GWAS (b) and two splice variants of *FSD3* (c). Average gene expression levels were taken from six natural accessions (with 5 replicates per accession per treatment) and normalized relative to three reference genes (*SAND*, *YLS8*, and *TIP41-like*). Significance of differences was assessed using two-sample t-tests on all Ct values from both treatments per haplotype. Error bars represent SE. $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

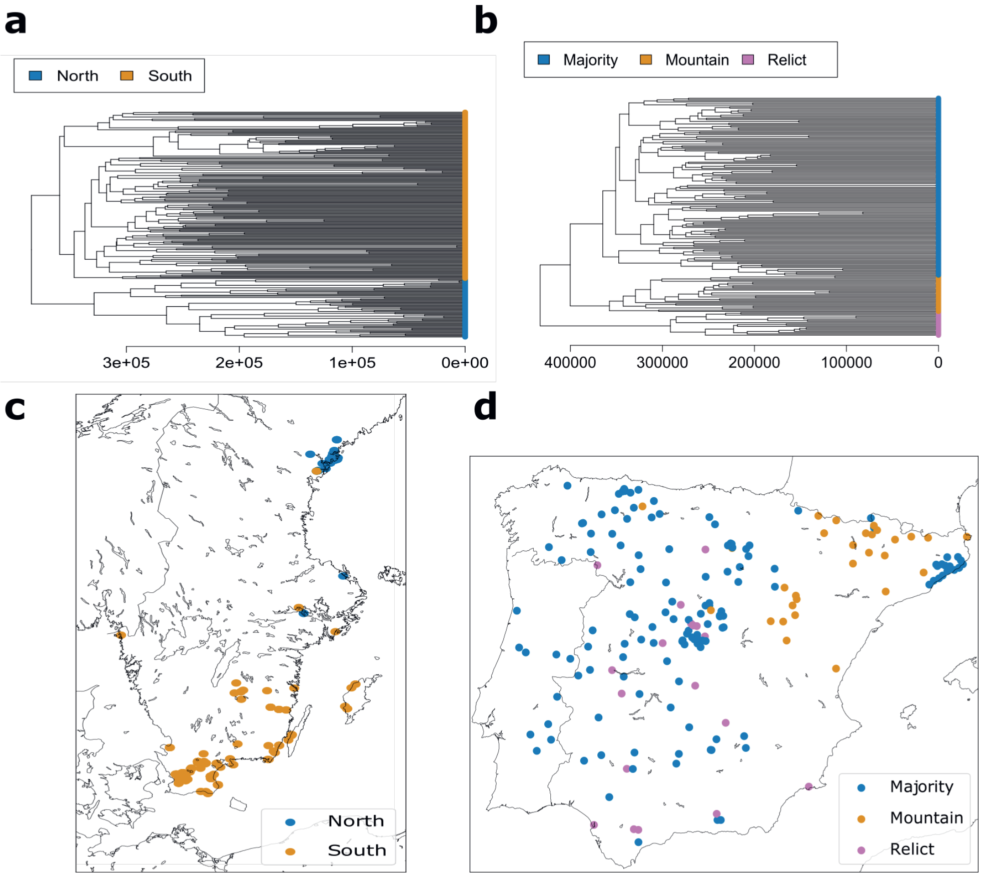


Figure S9: **Population structure of *A. thaliana* in Sweden and the Iberian peninsula.** (a) Dendrogram of Swedish accessions. (b) Dendrogram of Iberian accessions. (c) Geographical location of groups of Swedish accessions after cutting the dendrogram into two clusters. Groups correspond to accessions sampled in the north and south, as previously reported (Long *et al.*, 2013). (d) Geographical location of groups of Iberian accessions after cutting the dendrogram into three clusters. The two outgroups of that are distinct from the majority correspond to relicts and accessions sampled in mountains, as previously reported (Tabas-Madrid *et al.*, 2018).

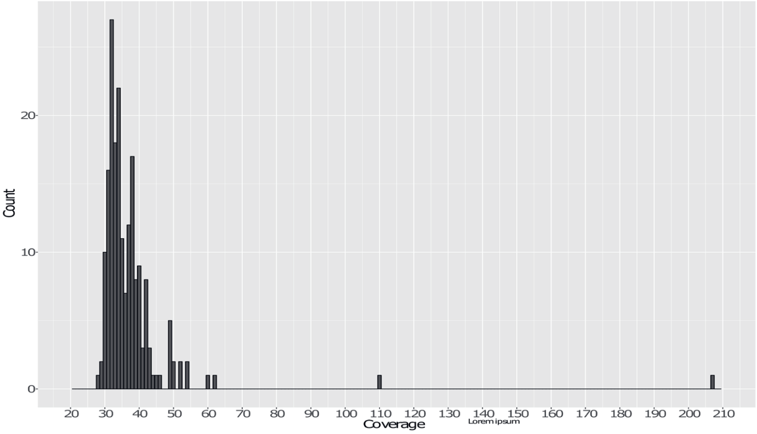


Figure S10: **Genomic coverage of each DartMap accession after pre-processing.** All samples have at least 28x coverage.

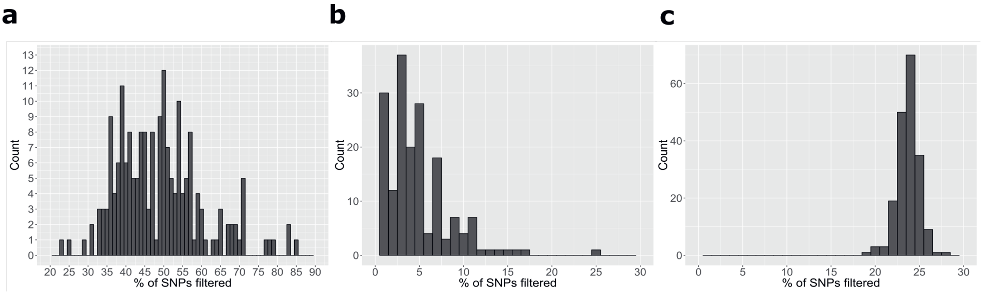


Figure S11: **Effect of filtering steps on number of retained SNPs for each sample.** The percentage of filtered SNPs is shown for the mitochondrial (a), chloroplast (b), and nuclear (c) callsets. The y-axes show the number of samples in which a particular percentage of SNPs (x-axes) were filtered.

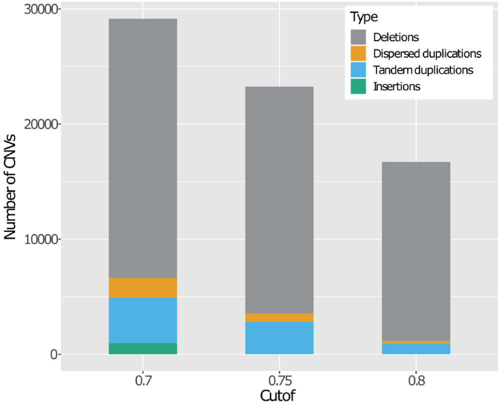


Figure S12: **Number and types of CNV detected at different cut-offs of Hecaton.** Insertions were not reported at high cut-offs, as Hecaton generally considers them to be inaccurate.

Chapter 5

The genome sequence of *Hirschfeldia incana*, a new model to improve photosynthetic light-use efficiency

Francesco Garassino[1,*], Raúl Y. Wijfjes[2,5,*], René Boesten[1,*], Francisca Reyes Marquez[1], Frank F. M. Becker[1], Vittoria Clapero[1,6], Iris van den Hatert[2], Rens Holmer[2], M. Eric Schranz[3], Jeremy Harbinson[4], Dick de Ridder[2], Sandra Smit[2,†], and Mark G. M. Aarts[1,†]

[1] Laboratory of Genetics, Wageningen University & Research
[2] Bioinformatics Group, Wageningen University & Research
[3] Biosystematics Group, Wageningen University & Research
[4] Laboratory of Biophysics, Wageningen University & Research
[5] Current affiliation: Faculty of Biology, Ludwig Maximilian University of Munich
[6] Current affiliation: Max Planck Institute for Molecular Plant Physiology
[*] These authors contributed equally to this work
[†] Corresponding authors: Sandra Smit, sandra.smit@wur.nl;
Mark Aarts, mark.aarts@wur.nl

Published in The Plant Journal: https://doi.org/10.1111/tpj.16005

Summary

Photosynthesis is a key process in sustaining plant and human life. Improving the photosynthetic capacity of agricultural crops is an attractive means to increase their yields. While the core mechanisms of photosynthesis are highly conserved in C_3 plants, these mechanisms are very flexible, allowing considerable diversity in photosynthetic properties. Amongst this diversity is the maintenance of high photosynthetic light-use efficiency at high irradiance as identified in a small number of exceptional C_3 species. *Hirschfeldia incana*, a member of the Brassicaceae family, is such an exceptional species, and because it is easy to grow, is an excellent model for studying the genetic and physiological basis of this trait. Here, we present a reference genome of *H. incana* and confirm its high photosynthetic light-use efficiency. While *H. incana* has the highest photosynthetic rates found so far in the Brassicaceae, the light-saturated assimilation rates of closely related *Brassica rapa* and *Brassica nigra* are also high. The *H. incana* genome has extensively diversified from that of *B. rapa* and *B. nigra* through large chromosomal rearrangements, species-specific transposon activity, and differential retention of duplicated genes. Duplicated genes in *H. incana*, *B. rapa* and *B. nigra* that are involved in photosynthesis and/or photoprotection show a positive correlation between copy number and gene expression, providing leads into the mechanisms underlying the high photosynthetic efficiency of these species. Our work demonstrates that the *H. incana* genome serves as a valuable resource for studying the evolution of high photosynthetic light-use efficiency and enhancing photosynthetic rates in crop species.

Introduction

Photosynthesis is the biophysical and biochemical process that sustains most life on planet Earth. The most common form of photosynthesis, oxygenic photosynthesis, uses solar energy to convert the inorganic carbon dioxide (CO_2) to organic carbon, typically represented as a carbohydrate, releasing molecular oxygen (O_2) from water in the process. Terrestrial plants provide by far the most conspicuous example of oxygenic photosynthesis (referred to as photosynthesis from now on for brevity) and are responsible for about 50% of the primary production of oxygen in the biosphere, with marine production by eukaryotic algae and cyanobacteria comprising the other 50%. Agriculture depends on primary production by plants, so expanding our knowledge of photosynthesis is crucial if we are to meet many of the pressing global challenges faced by mankind.

One of these challenges is the need to substantially increase the yield of agricultural crops to meet the increasing demand not only for food and fodder, but also for fibers and similar plant products, and organic precursors for the chemical industry as it transitions away from fossil carbon sources. A major yield-related trait is the conversion efficiency of absorbed solar irradiance to biomass (ε_c (S. P. Long *et al*., 2006)), a parameter which is strongly influenced by the light-use efficiency of photosynthesis. As light intensity, or irradiance, increases, the photosynthetic light-use efficiency of leaves and other photosynthetic organs decreases, which leads ultimately to the light-saturation of photosynthesis (Genty & Harbinson, 1996; Gitelson *et al*., 2015; J. Gu *et al*., 2017; Monneveux *et al*., 2003; Murchie *et al*., 1999; Turner *et al*., 2003). Once light-saturation is reached, any additional light will not lead to a further increase in the photosynthetic rate and may even be detrimental to photosynthesis. The threshold for light saturation generally lies far below the maximum level of irradiance experienced in the field or greenhouse (X.-G. Zhu *et al*., 2010) and for most C_3 crops this light-saturation phenomenon is an aspect of their photosynthesis which remains to be increased in order to increase yield. Improving the photosynthetic light-use efficiency of crop plants thus paves the way towards increasing their ε_c and ultimately their yield (Flood *et al*., 2011; Furbank *et al*., 2019; Lawson *et al*., 2012; von Caemmerer & Evans, 2010; X.-G. Zhu *et al*., 2010), as recently shown in soybean (De Souza *et al*., 2022).

The means with which to reduce the loss of photosynthetic light-use efficiency in crop plants with increasing irradiance already exists in nature. Most temperate zone crop species, alongside tropical crops species like rice, make use of the C_3 photosynthetic pathway, which is the original and ancestral photosynthetic pathway in higher plants, with the alternative CAM and C_4 pathways having evolved as an adaptation to heat and drought, and low CO_2 levels. Due to several issues associated with the C_3 pathway compared to the C_4 pathway, the maximum photosynthesis rates commonly observed among C_3 species are generally lower than those of C_4 ones. Although the core mechanisms of photosynthesis are highly conserved (Leister, 2019; Shi *et al*., 2005), natural variation in photosynthesis rates has been observed for major crops such as wheat (Driever *et al*., 2014), rice (J. Gu *et al*., 2012, 2014), maize (Strigens *et al*., 2013), soybean (Gilbert *et al*., 2011), sorghum (Ortiz *et al*., 2017), as well as for the model species *Arabidopsis thaliana* (van Rooijen *et al*., 2015; Van Rooijen *et al*., 2017). Much higher photosynthesis rates can be expected in species that are more ecologically specialized (van Bezouw *et al*., 2019). Exceptionally high light-use efficiencies (and high assimilation rates) at high irradiance has been found previously in species growing in the Sonoran Desert, such as *Amaranthus palmeri*, *Chylismia claviformis*, *Eremalche rotundifolia*, and *Palafoxia linearis* (Ehleringer, 1985; Werk *et al*., 1983). Although data collected on these species provided clues about the anatomical and physiological basis of their high photosynthesis rates (Gibson, 1998; Werk *et al*., 1983), a comprehensive ecophysiological explanation of their phenotypes is still missing.

To understand the physiological and genetic basis of this more efficient photosynthesis at high irradiance, a suitable model species is needed. To date, of the handful of species showing high light-use efficiency that have been described (Ehleringer, 1985; Werk *et al*., 1983), none would qualify as a model species due to a combination of complex genetics and difficulties in growing in laboratory conditions (e.g. difficult seed germination). Taking inspiration from *A. thaliana*, an attractive model species for high light-use efficiency would need to be easily grown in either regular irradiance (typically up to 600 μmol $m^{-2} \cdot s^{-1}$) and high-light laboratory conditions; have a high-quality reference genome; be a diploid species capable of producing a large number of progeny (hundreds of seeds from a single mother plant) with a short generation time; germinate easily and have easily stored seed; and allow for both inbreeding and outcrossing (Koornneef & Meinke, 2010; Somerville & Koornneef, 2002).

Hirschfeldia incana (L.) Lagr.-Foss. is an excellent candidate that fulfils these requirements. *H. incana* is a thermophilous and nitrophilous annual species native to the Mediterranean basin and the Middle-East, but currently widespread in most warm-temperate regions of the world (Kole, 2011). It is generally self-incompatible and thus allogamous, but a degree of self-compatibility has been observed in natural populations (P. L. M. Lee *et al*., 2004). Although it makes use of the C_3 pathway, *H. incana* has a very high photosynthesis rate at high irradiance (Canvin *et al*., 1980), much higher than that of the C_3 crop species wheat (Driever *et al*., 2014) and rice (J. Gu *et al*., 2012), more in the range of C_4 species (Crafts-Brandner & Salvucci, 2002; Leakey *et al*., 2006). Besides its exceptional physiological properties, *H. incana* is also an attractive model species for practical and genetic reasons. It shows fast and sustained growth in laboratory conditions and is a member of the Brassiceae tribe within the well-studied Brassicaceae family, allowing the use of many genetic and genomic resources developed for the model species *A. thaliana* and its close relatives *Brassica rapa* (Belser *et al*., 2018; Choi *et al*., 2007; *Brassica rapa* Genome Sequencing Project Consortium *et al*., 2011; Kim *et al*., 2009; Zhang *et al*., 2018), *Brassica nigra* (Paritosh *et al*., 2020; Perumal *et al*., 2020), *Brassica oleracea* (Belser *et al*., 2018; S. Liu *et al*., 2014; X. Wang *et al*., 2011), and *Brassica napus* (Bancroft *et al*., 2011; Chalhoub *et al*., 2014). Yet, *H. incana* has received little attention from the research community so far, being recognised mainly as a possible lead (Pb) hyperaccumulator

(Auguy *et al*., 2013, 2016; Fahr *et al*., 2015) and for the ecological implications of its occurrence as a weed (Darmency & Fleury, 2000; P. L. M. Lee *et al*., 2004; Y. Liu *et al*., 2013; Mira *et al*., 2019; Sánchez-Yélamo, 2009).

Here we present a genomic assembly and gene set of *H. incana*. We expect these data to lay the foundation for studying photosynthetic light-use efficiency and improving this trait in crop species, through a process of candidate gene identification followed by phenotypic validation using genetic modification and/or gene editing. First, we directly compare the photosynthetic rate of *H. incana* at high irradiance to that of the Brassicaceae species *B. rapa*, *B. nigra*, and *A. thaliana* to affirm its high light-use efficiency. Second, we characterize how the *H. incana* genome differs from that of other members of the Brassicaceae family, specifically focusing on differences in numbers of gene copies. Finally, we report on whether such differences translate to differential expression of genes expected to mediate high light-use efficiency. Our work demonstrates how the genome assembly of *H. incana* serves as a valuable resource to elucidate the genetic basis of high photosynthetic performance and for studying the evolution of this trait in the Brassicaceae family.

Results

Hirschfeldia incana has an exceptionally high rate of photosynthesis

High photosynthesis rates have been reported for *H. incana* in 1980 (Canvin *et al*., 1980). We performed new measurements in order to compare the performance of *H. incana* with that of close relatives and the well-established model species *Arabidopsis thaliana* (Figure 1, Table S1). Gross CO_2 assimilation rates differed significantly between these species (Table S2). The two *H. incana* accessions had the highest average gross CO_2 assimilation rates above an irradiance of 550 µmol $m^{-2} \cdot s^{-1}$ (PAR), although only ‘Burgos’ had a statistically significant higher rate than the other species (Table S3). Gross assimilation rates are independent of CO_2 release by mitochondrial respiration and therefore a better indication of photosynthetic capacity than net photosynthesis rates, however, net photosynthesis rates showed a similar trend, but with larger differences between the two *H. incana* genotypes (Figure S1, Table S3), attributed to differences in rates of daytime dark respiration (Rd, Table S4).

A reference genome of *H. incana*

We assembled a scaffold-level reference genome of *H. incana* based on one genotype of the ‘Nijmegen’ accession. *H. incana* is a predominantly self-incompatible species, but ‘Nijmegen’ produces, nonetheless, some 100 seeds per plant upon self-pollination, which is much more than ‘Burgos’, so inbreeding is possible. Inbred genotypes are expected to be much more homozygous than open-pollinated genotypes, which is preferred for genome sequencing. Therefore, the ‘Nijmegen’ accession was inbred for six generations, prior to whole-genome sequencing. Its haploid genome size was estimated to be 487 Mb, based on flow cytometry (Table S5). This estimate is smaller than the previously reported genome size estimates of

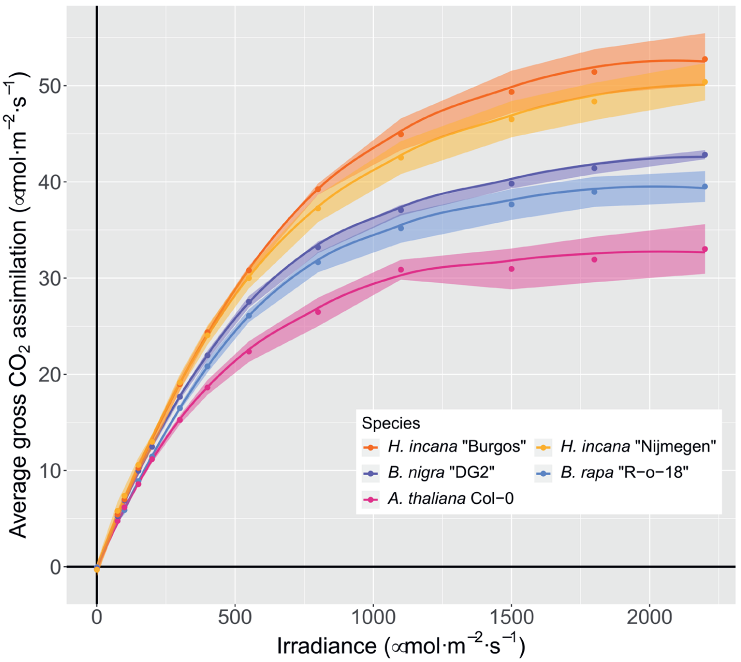


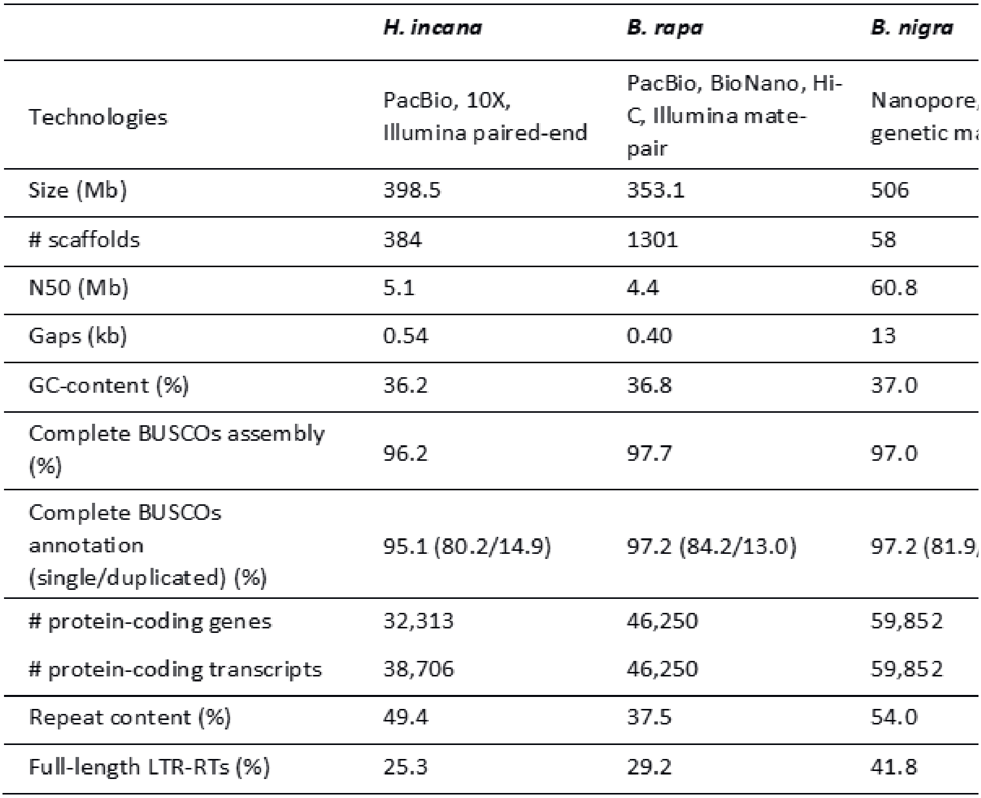
Figure 1: **Two *H. incana* genotypes have a higher gross CO_2 assimilation at high irradiance than genotypes of close relatives.** Light-response curves for *H. incana*, *B. rapa*, *B. nigra*, and *A. thaliana* accessions adapted to high levels of irradiance. Each point represents the mean gross CO_2 assimilation value of three (*B. rapa*) or four leaves coming from independent plants. Ribbons represent the standard error of the means. The lines indicate trends in gross assimilation for the various species and were obtained via LOESS smoothing.

B. rapa (529 Mb) and *B. nigra* (632 Mb) (Johnston *et al*., 2005). Chromosome counts from root tip squashes showed seven pairs of chromosomes (2n=14) (Figure S2), consistent with previous reports (Anderson & Warwick, 1999; Kole, 2011).

We generated DNA sequencing data consisting of 56 Gb of PacBio long reads (115-fold genome coverage, based on the genome size estimate), 46 Gb of 10X Genomics synthetic long reads (94-fold coverage, referred to as “10X” from now on for brevity), and 33 Gb of Illumina paired-end short reads (68-fold coverage). In addition, we generated 7.5 Gb of RNA sequencing (RNA-seq) data from leaf tissue for annotation purposes. Summary statistics and accession numbers can be found in Table S6. A k-mer analysis of Illumina data resulted in a haploid genome size estimate of 325 Mb, with a low level of heterozygosity (1.2%).

Using a hybrid assembly strategy, we produced a nuclear genome assembly of 399 Mb of sequence in 384 scaffolds with an N50 of 5.1 Mb (Table 1, see Table S7 for the full report generated by QUAST (Gurevich *et al*., 2013)). The assembly size is slightly larger than the genome size estimated from Illumina read k-mers (325 Mb), but smaller

Table 1: **Genomic properties of assemblies generated of *H. incana* 'Nijmegen' (this study), *B. rapa* Chiifu 401-42 (Zhang *et al*., 2018) and *B. nigra* Ni100 (Perumal *et al*., 2020).**



	H. incana	***B. rapa***	***B. nigra***
Technologies	PacBio, 10X, Illumina paired-end	PacBio, BioNano, Hi-C, Illumina mate-pair	Nanopore, genetic m
Size (Mb)	398.5	353.1	506
# scaffolds	384	1301	58
N50 (Mb)	5.1	4.4	60.8
Gaps (kb)	0.54	0.40	13
GC-content (%)	36.2	36.8	37.0
Complete BUSCOs assembly (%)	96.2	97.7	97.0
Complete BUSCOs annotation (single/duplicated) (%)	95.1 (80.2/14.9)	97.2 (84.2/13.0)	97.2 (81.9
# protein-coding genes	32,313	46,250	59,852
# protein-coding transcripts	38,706	46,250	59,852
Repeat content (%)	49.4	37.5	54.0
Full-length LTR-RTs (%)	25.3	29.2	41.8

than the typical overestimate (Sun *et al*., 2018) based on flow cytometry (487 Mb). Besides the nuclear genome, we assembled the mitochondrial and chloroplast genomes of *H. incana* into single sequences of 253 and 153 kb, and annotated the latter. The chloroplast assembly is typical for a Brassicaceae species, as it is nearly identical to chloroplast assemblies of *A. thaliana*, *B. rapa*, and *B. nigra* in terms of length and number of annotated genes (Table S8), and thus very useful for phylogenetic comparison of *H. incana* with other Brassicaceae.

The nuclear genome assembly is near-complete and structurally consistent with the underlying read data of *H. incana* 'Nijmegen' (Table S9). The high mapping rate of Illumina and 10X reads (>93%) suggest completeness, while the lower mapping rate of PacBio reads (81.5%) suggests some misassemblies or missing regions, likely repeats. The high mapping rate of RNA-seq reads (93.6%) also shows the gene space is near complete. We estimated the base-level error rate of the assembly to be 1 per 50 kb at most, based on variant calling using the mapped reads, resulting in 8,374 and 4,166 homozygous variants from the Illumina and 10X read alignments respectively.

We have annotated 32,313 gene models and 38,706 transcripts in the *H. incana* assembly (Table 1). This is a conservative annotation, based on filtering 64,546 initial gene models resulting from ab initio protein alignment, and RNA-seq based predictions. Our filtering approach is more stringent than those used to generate the *B. rapa* and *B. nigra* annotations, which explains why we report a lower number of genes and transcripts for *H. incana* (Table 1) than for both *Brassica* species. The annotation is expected to cover the large majority of the *H. incana* gene space. It contains 95.1% of 1,440 single-copy orthologs (BUSCOs) conserved in the Embryophyta plant clade, comparable to the percentages found for *B. rapa* and *B. nigra* (both 97.2%) (Table 1). The ratio of single to multiple copies is similar to that of *B. rapa* and *B. nigra* (Table 1), suggesting that the 14.9% of the BUSCOs present in multiple copies are true gene duplications shared by several species of the Brassiceae tribe. We additionally evaluated the completeness of the annotation by aligning protein sequences of *B. rapa* to the assembly and determining overlap between protein alignments and annotated genes. 30,552 out of the 37,387 protein alignments (81.7%) corroborate the annotation, as they completely or partially overlap with an annotated protein-coding gene. 2,570 (6.9%) of the protein alignments completely or partially overlap with an annotated repeat, suggesting that the aligned *B. rapa* proteins correspond to transposable elements. The remainder of the *B. rapa* proteins completely or partially overlap with gene models that were filtered (3,945 or 10.6%) or do not overlap with any annotated element at all (320 or 0.9%), indicating a small number of genes that are potentially missing from the annotation. Based on these observations, we conclude that the *H. incana* assembly is mostly contiguous, correct, and complete, making it a solid foundation for comparative analyses with other Brassicaceae.

The genome of *H. incana* extensively diversified from that of *B. rapa* and *B. nigra*

We utilized our assembly to explore the genomic divergence between *H. incana*, *B. rapa*, and *B. nigra*, all members of the same Brassiceae tribe. A substantial degree of divergence is expected between the three species due to different processes of post-polyploid diploidization, i.e. the process in which polyploid genomes get extensively rearranged as they return to a diploid state (Mandáková & Lysak, 2018), following the ancient two-step genome triplication event shared by all Brassiceae (*Brassica rapa* Genome Sequencing Project Consortium *et al*., 2011; He *et al*., 2021; Lysak *et al*., 2005). Part of this divergence may have facilitated the evolution of the exceptionally high rate of photosynthesis at high irradiance in *H. incana*.

We first assessed the phylogenetic relationship between *H. incana*, *B. rapa*, and *B. nigra* by constructing phylogenetic trees based on homologous nuclear and chloroplast genes, using *A. thaliana* as the outgroup. Both trees are congruent with each other and suggest that *H. incana* is more closely related to *B. nigra* than *B. rapa* (Figures 2a and 2b). This is corroborated by the median rate of synonymous substitutions between the syntenic orthologs (Ks) of the three species, which correspond to speciation events with an estimated time of 10.4 (*H. incana* - *B. nigra*) and 11.6 (*H. incana* - *B. rapa*) million years ago (mya) (Figure 2c) as obtained by dividing the median Ks of each curve by the rate of 8.22 × 10^{-9} synonymous substitutions per year established for Brassicaceae species (Beilstein *et al*., 2010). Our results are consistent with a previous phylogenetic analysis based on four intergenic chloroplast regions (Arias & Pires, 2012), but slightly differ from a more recently constructed phylogeny of the Brassicaceae, which was only based on 113 nuclear genes (C.-H. Huang *et al*., 2016), while we consider many more.

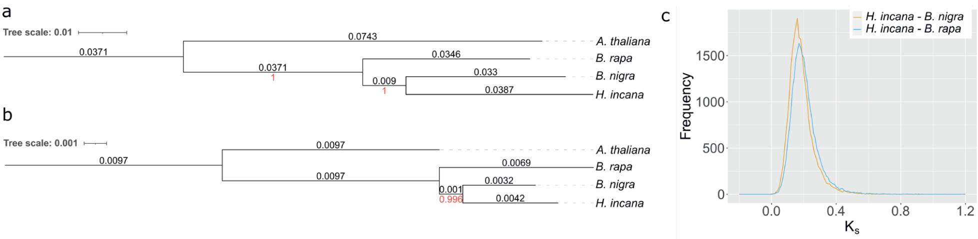


Figure 2: ***Hirschfeldia incana* is more closely related to *Brassica nigra* than *Brassica rapa*.** (a-b) Phylogenetic trees of *H. incana*, *B. rapa*, *B. nigra*, and *A. thaliana* (outgroup), based on nuclear (a) and chloroplast (b) genes. Branch lengths (black) and bootstrap values (red) are displayed above and below each branch, respectively. (c) Distributions of the rates of synonymous substitutions between 25,127 and 26,137 orthologous gene pairs of *H. incana* - *B. rapa* and *H. incana* - *B. nigra*, respectively. Both distributions show a single peak corresponding to speciation events with an estimated time of 11.6 (*H. incana* - *B. rapa*) and 10.4 (*H. incana* - *B. nigra*) million years ago (mya).

We determined rearrangements between the genomes of *H. incana* - *B. rapa* and *H. incana* - *B. nigra* by comparing the order of syntenic orthologs between their assemblies. On a small scale, most genomic regions of *H. incana* are syntenic (not rearranged) with *B. rapa* and *B. nigra*, as 77.7% and 81.0% of the genes of *H. incana* could be clustered in collinear blocks containing a minimum of four orthologous pairs of *H. incana* - *B. rapa* and *H. incana* - *B. nigra*, respectively. Gene order is less conserved when comparing larger blocks, indicating several rearrangements between the twenty largest scaffolds of *H. incana* (covering 43.6% of the assembly) and the chromosomes of the other two species (Figures 3a and S3). For example, the two largest scaffolds of the *H. incana* assembly both contain inversions and/or translocations relative to their homologous chromosomes in *B. rapa* and *B. nigra*. A similar pattern of rearrangements of small collinear blocks was observed between the genomes of *B. rapa* and *B. nigra* in previous work (Z. He *et al*., 2021).

We further examined genomic differentiation between the three species by comparing their transposable element (TE) content. Of the assembly of *H. incana*, 49.4% consists of repetitive elements (Table 1), of which most are long terminal repeat retrotransposons (LTR-RTs) (25.3% of the genome). These numbers are consistent with previous work that investigated the repeat content of the *H. incana* genome using genome skimming, and which reported a repeat content of 46.5% and LTR-RT content of 31.6% (Beric *et al*., 2021). We specifically focused our analyses on LTR-RTs, as LTR-RT expansion and contraction has been previously identified as a major driver of genomic differentiation between Brassiceae (Xu *et al*., 2018), even between different ecotypes of the same species (Cai *et al*., 2020). The composition of LTR-RTs in the *H. incana* assembly differs from that of the *B. rapa* and *B. nigra* assembly, as the majority of LTR-RTs consist of Gypsy elements in *H. incana*, consistent with earlier work (Beric *et al*., 2021), while Copia retrotransposons form the majority of LTR-RTs in the others (Figure 3b). Furthermore, the estimated insertion times of LTR-RTs vary between the three assemblies, as Gypsy and Copia elements in *H. incana* and *B. rapa* are predicted to have proliferated recently (< 1 mya) (Figure 3c-d), while Gypsy elements in *B. nigra* show a more varied distribution of insertion times (Figure 3c). A possible explanation of this shift could be that the *B. nigra* assembly was generated using longer reads than those used for the assemblies of *H. incana* and *B. rapa*, enabling it to capture a larger proportion of the centromeric regions, but we found no evidence that this introduced a bias towards longer insertion times of Gypsy elements (Figure 3c).

Taken together, the breakdown of genomic synteny and divergence of LTR-RT content indicate that the genome of *H. incana* extensively diversified from that of *B. rapa* and *B. nigra* following their shared genome triplication event.

Gene copy number expansion may contribute to high photosynthetic rates

Genomic differentiation can result in species-specific gains and losses of genes, and these may explain the differences in photosynthetic light-use efficiency between *H. incana*, *B. rapa*, and *B. nigra*. Given that the three species all share the same ancient genome triplication event (Z. He *et al*., 2021; Schranz *et al*., 2006), it is reasonable to assume that most differences originated through differential retention of duplicated genes, particularly those located in genomic blocks showing evidence of extensive fractionation since that event (Z. He *et al*., 2021). We investigated gene copy number variation between the three species, by clustering their annotated protein-coding genes with those of five other Brassicaceae species within (*Raphanus raphanistrum* and *Raphanus sativus*) and outside (*Aethionema arabicum*, *A. thaliana*, and *Sisymbrium irio*) the Brassiceae tribe into homology groups. The inclusion of *A. thaliana* allowed us to use its extensive genomic resources to functionally annotate the genes of the other species. The other four species were included to put the analysis into a broader phylogenetic context. *A. arabicum* is part of the Aethionema tribe which diverged from the core group of the Brassicaceae family, thus allowing us to identify highly conserved genes. *S. irio* is part of a different tribe than *A. thaliana* (Sisymbrieae) that is more closely related to the Brassiceae tribe (C.-H. Huang *et al*., 2016), but did not undergo the ancient genome triplication. *R. raphanistrum* and *R. sativus* are part of the *Raphanistrum* genus within the Brassiceae tribe and thus represent another set of species that underwent the genome triplication event shared by the whole tribe.

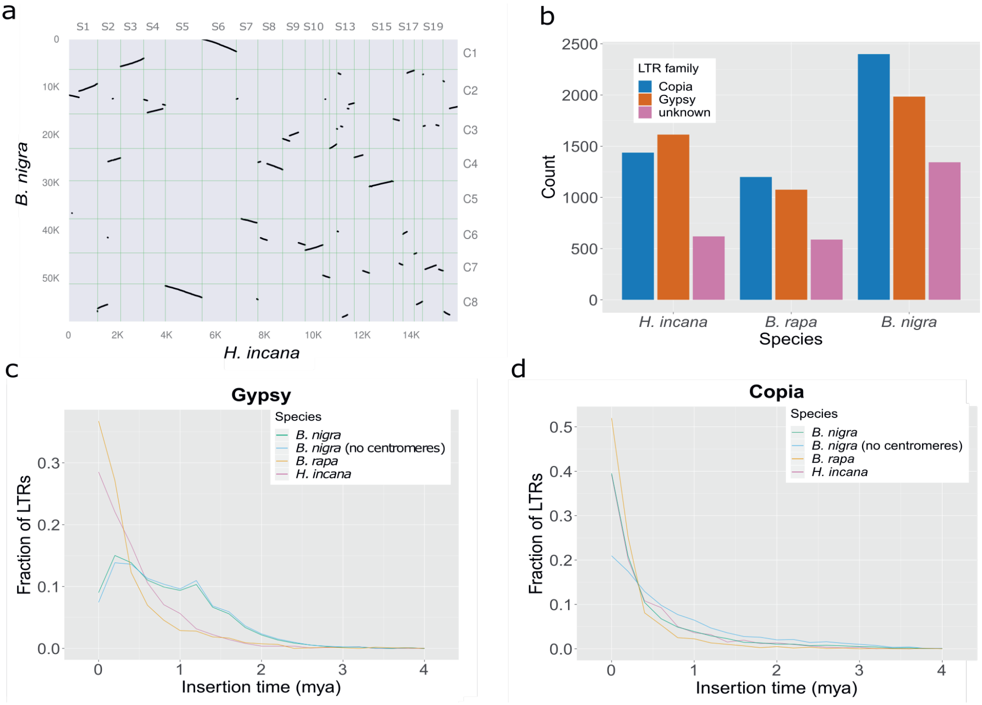


Figure 3: **The genome of *Hirschfeldia incana* extensively diversified from that of *Brassica rapa* and *Brassica nigra*.** (a) Orthologous syntenic blocks between the genomes of *H. incana* and *B. nigra*. Dots indicate pairs of syntenic orthologs that are found in the same order in both genomes according to sequence positions. Only the twenty largest scaffolds of *H. incana* (43.6% of the assembly) are shown for clarity. Axes labels correspond to the total number of genes annotated on the sequences (left and bottom) and identifiers of the scaffolds (top) or chromosomes (right). A dot plot visualizing orthologous syntenic blocks between *H. incana* and *B. rapa*, showing similar patterns, is found in Figure S3. (b) Frequency distribution of Long Terminal Repeat Retrotransposon (LTR-RT) families. LTR-RTs are classified as unknown if they contained elements of both Gypsy and Copia sequences and could thus not be reliably assigned to either of these families. (c) Frequency polygon (bin width = 0.2 mya) of the insertion times of Gypsy elements. (d) Frequency polygon (bin width = 0.2 mya) of insertion times of Copia elements.

Our analysis resulted in 20,331 groups containing at least one *H. incana* gene (Table S10). The composition of the homology groups agrees with the currently established phylogeny of the Brassicaceae (C.-H. Huang *et al*., 2016) as groups containing *H. incana* genes share the least number genes with *A. arabicum* (58.2%) and the greatest number of genes with species that are part of the Brassiceae tribe (86.3-95.6%). *H. incana* has a low fraction of species-specific homology groups (3.4%) compared to the seven other species, which can be attributed to the stringent filtering of the predicted gene models.

We focused on a subset of 15,097 groups containing at least one gene of *A. thaliana* and one of *H. incana*, as these could be extensively annotated through the transfer of Gene Ontology (GO) terms from *A. thaliana* genes to their respective groups. Consistent with the expectation that most genes quickly return to single-copy status following a whole genome duplication event (Z. Li *et al*., 2016), 70.2% of these groups contain a single gene of both *A. thaliana* and *H. incana*. Focusing on groups containing *A. thaliana* genes involved in photosynthesis (260 in total, Table S11), most contain a higher number of genes of *H. incana*, *B. rapa*, and *B. nigra*, compared to *A. thaliana* (Figure 4), consistent with the relatively higher photosynthetic light-use efficiency of *H. incana*, *B. rapa*, and *B. nigra* (Figure 1). That *H. incana* should have the highest efficiency of all the four species is not apparent from the gene copy number data because for most groups of genes, *H. incana* contains the same or a lower number of copies, relative to *B. rapa* and *B. nigra*. This is not a result of our conservative filtering approach, as we explicitly retained putative photosynthesis-related genes during our filtering procedure (Methods S1). Besides photosynthesis-related genes, we also analysed copy numbers of a more general set. 4,901 homology groups contain genes of which the copy number in *H. incana* is higher than in *A. thaliana*, and equal to or higher than those of *B. rapa* and *B. nigra* (Table S12). We estimate that 74.5 % of the duplicated gene pairs in *H. incana* (16,788 of the 22,535 analysed pairs) were duplicated through whole-genome duplication, 1.8% through tandem duplication, and the remaining 23.6% through another mode of duplication. Given that the increased photosynthetic light-use efficiency of *H. incana*, relative to *A. thaliana*, *B. rapa* and *B. nigra*, is particularly pronounced at high levels of irradiance (Figure 1), genes annotated with GO terms associated with photosynthesis and/or photoprotection are of particular interest. The 4,901 homology groups contain ample examples of such genes (Table S12), although the groups are not significantly enriched for any GO term specifically linked to photosynthesis and/or photoprotection (Table S13).

As gene copy number variation can considerably affect expression levels (Żmieńko *et al*., 2014), we hypothesized that retained copy number expansions of photosynthesis and photoprotection-associated genes in *H. incana*, *B. rapa* and *B. nigra* may aid the high photosynthetic capacities of these species (Figure 1). We therefore measured gene expression levels of nine genes for which there is inter-species copy number variation in two contrasting light conditions (200 μmol $m^{-2} \cdot s^{-1}$ and 1800 μmol $m^{-2} \cdot s^{-1}$), selecting genes with a function related to photosynthesis and/or photoprotection (Table 2). *A. thaliana*, the species with the lowest photosynthesis rates measured in this study, contains a single copy of each of the tested genes. For six genes, we observed a statistically significant positive correlation between gene expression level and gene copy number (Figure 5), with species showing higher or equal expression with an increasing number of copies (Figure S4). No such correlation was observed for the remaining three genes.

To test if the observed differences in gene expression are due to photosynthesis-related genes being more frequently more highly expressed

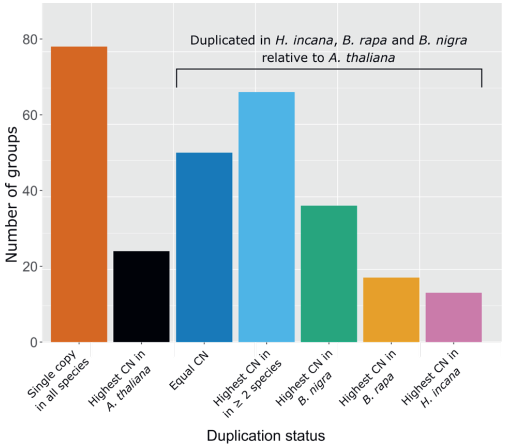


Figure 4: ***Hirschfeldia incana* retained fewer duplicated copies of photosynthesis-associated genes than *Brassica rapa* and *Brassica nigra*.** Bars show counts of homology groups containing genes associated with photosynthesis with different distributions of copy numbers (CN) in the four species (260 groups in total). For groups that contain a higher number of copies in *H. incana*, *B. rapa*, and *B. nigra* than in *A. thaliana*, it has been indicated whether the same number of copies is found in all three species (equal CN), whether there are two or more species that contain a higher number of copies than the other(s) (highest CN in ≥2 species), or whether there is a single species containing the highest number of copies.

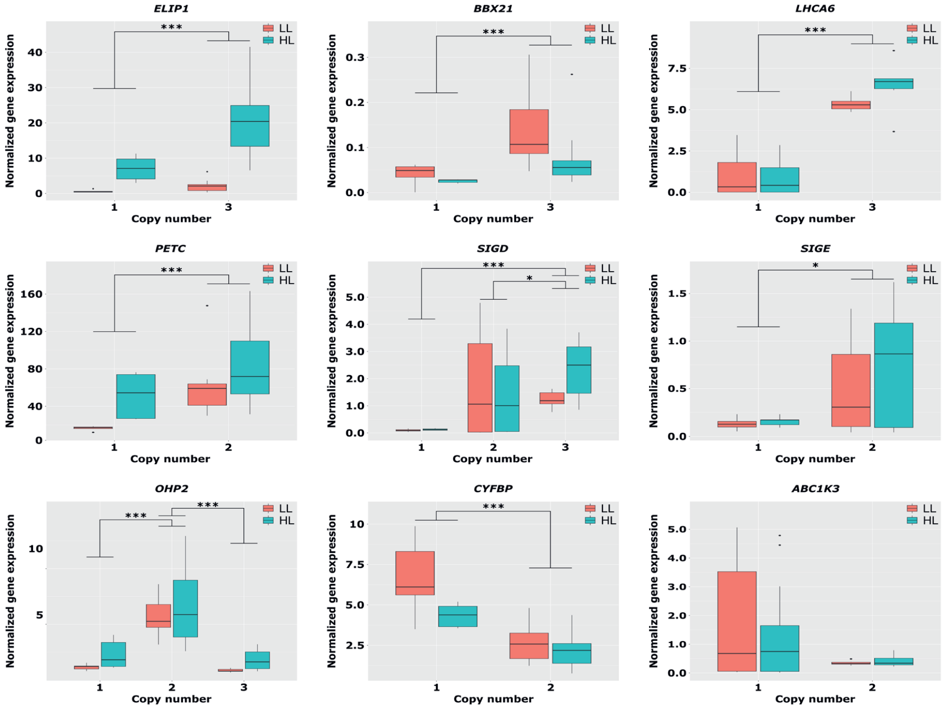
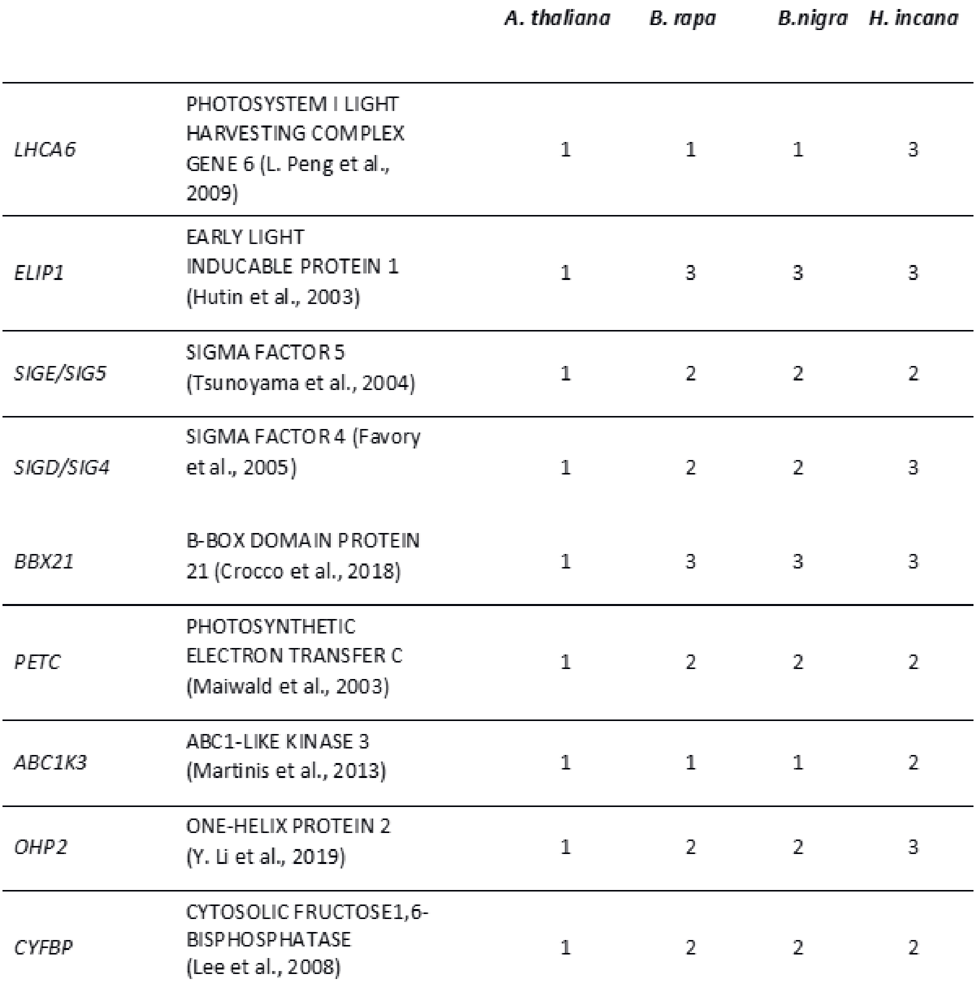


Figure 5: **Copy numbers of photosynthesis- and photoprotection-associated genes correlate with expression level.** Boxplots depict gene expression levels of *A. thaliana*, *B. rapa*, *B. nigra* and *H. incana* grown in 200 µmol m^{-2} · s^{-1} and 1500 µmol m^{-2} · s^{-1} . Gene expression levels were normalized against *H. incana* grown at 200 µmol m^{-2} · s^{-1} and subsequently grouped per gene copy number. Titles of graphs indicate gene names based on the *A. thaliana* gene nomenclature. $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

Table 2: **Genes with inter-specific copy number variation of which expression was measured.** All genes are annotated to function in photosynthesis and/or photoprotection.



		A. thaliana	*B. rapa*	*B.nigra*	*H. incana*
LHCA6	PHOTOSYSTEM I LIGHT HARVESTING COMPLEX GENE 6 (L. Peng et al., 2009)	1	1	1	3
ELIP1	EARLY LIGHT INDUCABLE PROTEIN 1 (Hutin et al., 2003)	1	3	3	3
SIGE/SIG5	SIGMA FACTOR 5 (Tsunoyama et al., 2004)	1	2	2	2
SIGD/SIG4	SIGMA FACTOR 4 (Favory et al., 2005)	1	2	2	3
BBX21	B-BOX DOMAIN PROTEIN 21 (Crocco et al., 2018)	1	3	3	3
PETC	PHOTOSYNTHETIC ELECTRON TRANSFER C (Maiwald et al., 2003)	1	2	2	2
ABC1K3	ABC1-LIKE KINASE 3 (Martinis et al., 2013)	1	1	1	2
OHP2	ONE-HELIX PROTEIN 2 (Y. Li et al., 2019)	1	2	2	3
CYFBP	CYTOSOLIC FRUCTOSE1,6-BISPHOSPHATASE (Lee et al., 2008)	1	2	2	2

in general in *B. rapa*, *B. nigra* and *H. incana*, compared to *A. thaliana*, we included nine additional genes in our experiment that are present in a single copy in all four species and involved in similar processes as the multi-copy genes. Although we find species-specific differences in expression for this set of genes, no consistently higher gene expression levels are found in *B. rapa*, *B. nigra* and *H. incana* compared to *A. thaliana* (Figure S5). Overall, our analyses suggest that the increased copy numbers of photosynthesis and photoprotection-associated genes in *H. incana*, *B. rapa* and *B. nigra*, relative to the *A. thaliana*, may contribute to their high photosynthetic efficiency, although this effect appears to not be specific to a particular species or level of irradiance.

Discussion

In this study, we generated a reference genome of *H. incana* to establish this species as a model for exceptional photosynthetic light-use efficiency at high irradiance. We find substantial differences in light-use efficiency, genomic structure, and gene content between *H. incana* and its close relatives. We discuss these results in terms of how they contributed to the evolution of the remarkable phenotype of *H. incana*.

Our results show an even higher photosynthetic light-use efficiency at high irradiance than previously reported for *H. incana* (Canvin *et al*., 1980), with photosynthesis rates varying marginally between both accessions. Examination of a wider set of *H. incana* accessions may identify genotypes with larger differences in photosynthesis rates, that would allow a quantitative genetic approach to identify alleles conferring high photosynthesis rates. Our measurements imply that the photosynthesis rates of this C_3 species are higher than those of the C_4 crop maize (Crafts-Brandner & Salvucci, 2002; Leakey *et al*., 2006) and almost two times higher than those typically reported from key cereal crop species with a C_3 photosynthetic metabolism, such as wheat (Driever *et al*., 2014) and rice (J. Gu *et al*., 2012), respectively. Furthermore, these rates are higher than those of closely related Brassicaceae species *B. rapa*, *B. nigra*, and the more distantly related *A. thaliana*. The photosynthesis rates we measured in *B. rapa* are also higher than previously reported (Pleban *et al*., 2018; Taylor *et al*., 2020). Although the rates presented in this study were obtained from plants grown in controlled, favourable conditions and thus could be an overestimation of rates in natural environments, the magnitude of the differences suggests that the *H. incana* genome holds essential information for the improvement of photosynthetic light-use efficiency in crops.

The reference genome of *H. incana* generated in this study provides the means to elucidate the genetic basis of this plant's exceptional rate of photosynthesis and how it evolved in this species. We estimate that *H. incana* diverged 11.6 and 10.4 mya from *B. rapa* and *B. nigra*, respectively, consistent with an earlier study that used a smaller set of nuclear genes (C.-H. Huang *et al*., 2016). These time points are close to the reported time at which *B. rapa* and *B. nigra* (Perumal *et al*., 2020) diverged from each other (11.5 mya) and the time at which the whole Brassicaceae family underwent a rapid radiation event (Franzke *et al*., 2009). This event may have been mediated by the expansion of grass-dominated ecosystems in the region inhabited by Brassicaceae family members at that time, which created new open habitats that favoured rapid diversification (Franzke *et al*., 2009). This expansion of grasslands is thought to have been driven by decreasing atmospheric CO_2 levels, and increasing aridity, which favoured the displacement of the then dominant C_3 plants by C_4 grasses (Edwards *et al*., 2010). We argue that climatic changes also drove the evolution of the high photosynthetic rates observed in *H. incana*; grassland, i.e. non-forested ecosystems may have provided the ephemeral niches with high irradiances that favoured the evolution of high photosynthetic rates. Species with high photosynthetic rates are currently found in Mediterranean and desert ecosystems (Ehleringer, 1985; Werk *et al*., 1983). The evolution of high rates of C_3 photosynthesis could therefore have paralleled the expansion of the C_4 photosynthesis pathway as an adaptation to low CO_2 levels and drought.

Our analyses suggest that the genome of *H. incana* extensively differentiated from the genomes of *B. rapa* and *B. nigra* since their time of divergence, through large genomic arrangements and differences in LTR-RT content. Previous analyses of natural *A. thaliana* accessions indicated that specific LTR-RT families show increased rates of proliferation in response to particular types of environmental stress (Baduel *et al*., 2021), which may explain the species-specific amplification of Gypsy elements that we observed in *H. incana*. Such elements may have been retained because this particular LTR-RT family generally inserts outside of exons (Baduel *et al*., 2021). We hypothesize that the differences in LTR-RT content between *H. incana*, *B. rapa*, and *B. nigra* were caused in part by Gypsy elements being less efficiently purged from the genome of *B. nigra* than from the others. An increased rate of LTR-RT removal, based on the ratio of solo LTRs to intact LTR-RTs, has also been observed in *B. rapa* relative to *B. oleracea* and it was speculated that this is caused by the increased rate of genetic recombination in the former (Zhao *et al*., 2013). Given that a similar negative correlation between local recombination rate and LTR-RT content was found in rice (Z. Tian *et al*., 2009), soybean (Du *et al*., 2012), and eukaryotes in general (Kent *et al*., 2017), the differences in predicted insertion times of Gypsy elements in *H. incana*, *B. rapa*, and *B. nigra* observed in this study may thus reflect different rates of genetic recombination in the three species. While it has been suggested that changes in recombination rate can be adaptive, there is little empirical evidence that supports this (Ritz *et al*., 2017). It would therefore be interesting to directly measure genome-wide rates of recombination of *H. incana*, *B. rapa*, and *B. nigra* and explore whether these are correlated with their rates of photosynthesis.

Further comparative analyses between the genomes of *H. incana*, *B. rapa*, *B. nigra*, and *A. thaliana* revealed numerous species-specific gains and losses of genes. For dosage-sensitive genes, such as those involved in transcriptional regulation, differences may not necessarily reflect adaptive selection. This category of genes was found to be consistently retained in multiple copies following polyploidy events across the Brassicaceae (Mandakova *et al*., 2017) and a wide group of angiosperms (Z. Li *et al*., 2016), which is hypothesized to be due to dosage constraints (Edger & Pires, 2009). Differences in copy number of such genes may thus reflect different rates of relaxation of dosage balance constraints and subsequent loss of duplicates through time, which is a neutral process.

On the other hand, there is reason to believe that gene duplications contributed to the evolution of the high light-use efficiency of *H. incana*. Gene duplications have been identified as important drivers of plant evolution and differences in gene copy number between species are often enriched for adaptive evolutionary traits (Dassanayake *et al*., 2011; Oh *et al*., 2013; Rizzon *et al*., 2006; Suryawanshi *et al*., 2016). Moreover, RT-qPCR analysis of nine duplicated genes associated with photosynthesis and/or photoprotection showed that the expression levels of six of them correlate with gene copy number. In contrast, nine photosynthetic genes present in a single copy in all species did not show significantly increased expression in *H. incana*, *B. nigra*, and *B. rapa* compared to *A. thaliana*, indicating that photosynthetic genes are not overexpressed in the former three species in general. This supports a putative role for gene duplications in mediating the high light-use efficiency achieved by *H. incana*, *B. nigra* and *B. rapa*.

The most striking genes of which copy number correlated with gene expression are *LHCA6* and *ELIP1*, involved in response to high light and having the highest expression in *H. incana* growing under high light (Figure S4). *LHCA6* encodes a light-harvesting complex I (LHCI) protein of photosystem I (PSI), that together with *LHCA5* is required to form a full-size NAD(P)H dehydrogenase (NDH)-PSI supercomplex (L. Peng *et al*., 2009). Higher expression of *LHCA6* might help the formation of the NDH-PSI complex, thought to help stabilise NDH under high irradiance conditions. In turn, NDH has proposed roles supporting the Calvin-Benson cycle's activity (Harbinson *et al*., 2022) and photoprotection by preventing overreduction at high light intensities (Munekage *et al*., 2004). *ELIP1* encodes for proteins with a proposed role in photoprotection, which is associated with high light stress (Heddad *et al*., 2006; Norén *et al*., 2003; Youssef *et al*., 2010). Increased expression of this gene is expected to make the photosynthetic apparatus of *H. incana* more resistant to photoinhibition at high levels of irradiance. While the *H. incana* genome harboured the highest number of copies of *LHCA6* when compared to the genomes of *A. thaliana*, *B. rapa*, and *B. nigra*, this is not the case for *ELIP1*, for which *H. incana*, *B. nigra* and *B. rapa* all have three copies as opposed to the single copy of *A. thaliana*. Therefore, although we can propose a role for gene duplications in achieving higher light-use efficiency, the exact nature of this role still remains unclear as it appears to not be completely dependent on species or light treatment.

While our gene expression analysis provides several promising leads, it only offers a glimpse of what may contribute to the high photosynthetic light-use efficiency of *H. incana*. Besides the nine genes included in this analysis, we identified many more genes with a high copy number in *H. incana* that warrant further investigations on a whole transcriptome level. Such investigations should not limit themselves to core photosynthetic genes, as *H. incana* can only attain high photosynthetic light-use efficiency through changes in many other traits that are outside the chloroplast, such as leaf architecture affecting mesophyll conductance to CO_2, the synthesis of carbohydrates in the cytosol, the transport of carbohydrates from the leaf, the uptake from the soil and the supply of nitrogen and other minerals to the leaf, the abundance and distribution of different leaf pigments, and (photo)respiration. Nor should they include duplicated genes only, as it is striking that *H. incana* shows a better high light-use efficiency than *B. rapa* and *B. nigra*, though it contains fewer photosynthesis-related genes than the latter two species. This points towards alternative scenarios in which adaptation of *H. incana* photosynthesis to high levels of irradiance occurred through regulation of expression of one copy of the photosynthesis-related genes, which relaxed selection on duplicate retention or even encouraged loss of duplicate copies, or through other traits, as described above.

To elucidate the exact genetic mechanisms underlying the high light-use efficiency of *H. incana*, a natural follow-up to this study is to perform comparative transcriptomic analyses of leaves of *H. incana*, *B. rapa*, and *B. nigra* under a range of different levels of irradiance and at different developmental stages. Genes that show copy number variation and are differentially expressed between *H. incana* and the latter two species, such as *LHCA6*, would then be prime candidates to further test for potential causality through e.g. knock-out mutant analysis. As previous work has shown that it is possible to cross distantly related Brassicaceae species (Katche *et al*., 2019), a useful approach to further pinpoint the causal genes is to establish a genetic mapping population between *H. incana* and a Brassicaceae species with regular light-use efficiency and perform quantitative trait locus analyses of photosynthetic traits segregating within the population. It would also be useful to expand comparative genome and transcriptome analyses to plant species outside of the Brassicaceae family that show high photosynthetic light-use efficiency, such as the aforementioned *A. palmeri*, *C. claviformis*, *E. rotundifolia*, and *P. linearis*. Such expanded analyses could be informative, for instance to investigate amino acid substitutions or lateral gene transfer specific to species with high photosynthetic light-use efficiency. Furthermore, transcriptome data may indicate genes showing differences in expression between such species and those that are less efficient, providing further insight into which genes contribute to the evolution of this trait.

Conclusions

H. incana has an exceptional rate of photosynthesis at high irradiance. We generated a near-complete reference genome of this species and found evidence suggesting that its exceptional rate evolved through differential retention of duplicated genes. Taken together, our work provides several promising leads that may explain the high photosynthetic light-use efficiency of *H. incana* and we expect the reference genome generated in this study to be a valuable resource for improving this efficiency in crop cultivars.

Materials and Methods

Plant material

Hirschfeldia incana accessions 'Nijmegen' and 'Burgos' were used. 'Nijmegen' is an inbred line (> six rounds of inbreeding) originally collected in Nijmegen, The Netherlands. Seeds of 'Burgos' were originally collected near Burgos, Spain. Furthermore, *Brassica nigra* accession 'DG2', sampled from a natural population near Wageningen, The Netherlands, the *Brassica rapa* inbred line 'R-o-18' (Bagheri *et al*., 2012; Stephenson *et al*., 2010), and the *Arabidopsis thaliana* Col-0 accession were used.

Measurements of photosynthesis rates

Seeds of *H. incana* 'Nijmegen', *H. incana* 'Burgos', *B. rapa* 'R-o-18', *B. nigra* 'DG2', and *A. thaliana* Col-0 were sown in 3-L pots filled with a peat-based potting mixture. Plants were grown in a climate chamber with a photoperiod of 12 hours and day and night temperatures of 23 and 20 °C, respectively. Humidity and CO_2 levels were set at 70% and 400 ppm. The chamber was equipped with high-output LED light modules (VYPR2p, Fluence by OSRAM). Plants were watered daily with a custom nutrient solution (0.6 mM NH_4^+, 3.6 mM K^+, 2 mM Ca^{2+}, 0.91 mM Mg^{2+}, 6.2 mM NO_3^-, 1.66 mM SO_4^{2-}, 0.5 mM P, 35 µM Fe^{3+}, 8 µM Mn^{2+}, 5 µM Zn^{2+}, 20 µM B, 0.5 µM Cu^{2+}, 0.5 µM Mo^{4+}). The seeds were germinated at an irradiance of 300 µmol $m^{-2} \cdot s^{-1}$, and the same irradiance was maintained to let seedlings establish. On day 14, 21, and 25 after sowing, the irradiance was raised to 600, 1200, and 1800 µmol $m^{-2} \cdot s^{-1}$, respectively.

The photosynthetic metabolism of young, fully expanded leaves developed under 1800 µmol $\cdot m^{-2} \cdot s^{-1}$ of light was measured with a LI-COR 6400xt portable gas exchange system (LI-COR Biosciences) equipped with a 2 cm^2 fluorescence chamber head. "Rapid" descending light-response curves were measured between 30 and 35 days after sowing to accommodate differences in growth rates of the different species on one leaf from four *H. incana* 'Nijmegen', *H. incana* 'Burgos', *B. nigra* 'DG2', and *A. thaliana* Col-0 plants, and three *B. rapa* 'R-o-18' plants. The net assimilation rates of the plants were measured at thirteen different levels of irradiance ranging from 0 to 2200 µmol $\cdot m^{-2} \cdot s^{-1}$. During measurements, leaf temperature was kept constant at 25 °C and reference CO^2 concentration was kept at 400 µmol mol^{-1}. Water in the reference air flux was regulated in order to achieve vapour-pressure deficit values comprised between 0.8 and 1.2 kPa. Light response curve parameters (Amax: net CO_2 assimilation at saturating irradiance, φ: apparent quantum yield of CO_2 assimilation, Rd: daytime dark respiration rate, and θ: curve convexity) were estimated for each species through non-linear least squares regression of a non-rectangular hyperbola (Marshall & Biscoe, 1980) with the R package "photosynthesis" (version 2.0.0) (Stinziano *et al*., 2020). An indication of gross assimilation rates for each species was subsequently generated by adding the daytime dark respiration rate (Rd) estimated for each species to the species' net assimilation rates.

Differences in net and gross assimilation rates were tested at each light level of the light-response curve with a one-way ANOVA on the "genotype" experimental factor. Pairwise comparisons between the assimilation rates of the different genotypes at each light level were subsequently performed and tested with the Tukey-Kramer extension of Tukey's range test. The *p*-value threshold for statistical significance was set at α = 0.05.

Flow cytometry

Leaf samples of the *H. incana* genotypes 'Burgos' and 'Nijmegen' and *A. thaliana* Col-0 were analysed for nuclear DNA content by flow cytometry (Plant Cytometry Services B.V., Didam, the Netherlands). Seven, three and five biological replicates were measured over separate rounds of analysis for *H. incana* 'Nijmegen' *H. incana* 'Burgos', and *A. thaliana* Col-0, respectively. Nuclei were extracted from leaf samples following the method by (Arumuganathan & Earle, 1991), and stained with

4′,6-diamidino-2-phenylindole (DAPI). The DNA content of nuclei relative to that of the reference species *Monstera deliciosa* was determined on a CyFlow Ploidy Analyser machine (Sysmex Corporation, Kobe, Japan). A haploid flow cytometry estimate of 157 Mb was used for *A. thaliana*, resulting from comparisons of nuclear DNA content of this species and other model organisms (Bennett *et al*., 2003). Haploid genome size estimates for the *H. incana* genotypes were obtained by multiplying the *H. incana*-to-*M. deliciosa* ratio by the haploid *A. thaliana* estimate and dividing this product by the average *A. thaliana*-to-*M. deliciosa* ratio.

Chromosome counting

Root tips (approximately 1 cm long) were collected from young, fast-growing rootlets of multiple *H. incana* 'Nijmegen' plants and pre-treated for 3 h at room temperature with a 0.2 mM 8-hydroxyquinoline solution. After pre-treatment, the 8-hydroxyquinoline solution was replaced with freshly prepared Carnoy fixative (1:3 (v/v) acetic acid - ethanol solution) and maintained at room temperature for half a day. Root tips were then rinsed with 70% ethanol for three times to remove remaining fixative and stored in 70% ethanol at 4 °C until further use. Prior to slide preparation, root tips were rinsed twice in Milli-Q (MQ) water before adding 1:1 solution of a pectolytic enzymatic digestion solution (1% Cellulase from *Trichoderma*, 1% Cytohelicase from *Helix Promatia*, 1% Pecolyase from *Aspergillus japonicus*) and 10 M citric buffer. After one hour incubation at 37 °C, the enzymatic digestion solution was replaced by MQ water. The digested root tips were spread in 45% acetic acid over microscopy slides on a hot plate set at 45 °C, cells were fixed with freshly prepared Carnoy fixative, dried, and stained with 4′,6-diamidino-2-phenylindole (DAPI) dissolved in Vectashield mounting medium (Vector Laboratories Inc., Burlingame, US). Slides were imaged with an Axio Imager.Z2 fluorescence optical microscope coupled with an Axiocam 506 microscope camera (Carl Zeiss AG, Oberkochen, Germany) at 63x magnification. Chromosome numbers were counted in metaphase mitotic cells and averaged to obtain the reported number.

DNA and RNA isolation

Genomic DNA was extracted from *H. incana* 'Nijmegen' samples using a protocol modified from (S. Chang *et al*., 1993). The modifications consisted of adding 300 μL β-mercaptoethanol to the extraction buffer just before use. We added 0.7% isopropanol to the supernatant instead of 10 M LiCl and then divided the total volume into 1 mL aliquots for subsequent extractions. The pellet was dissolved in 500 μL of SSTE which was preheated to 50 °C before use. The final pellets were dissolved in 50 μL MQ water and then pooled at the end of the extraction process. DNA used for Illumina and 10X Genomics sequencing was extracted from flower material, while leaf material was used for the PacBio sequencing, all originating from the same plant.

Total RNA was extracted from leaf material of *H. incana* 'Nijmegen' from a different plant than the one used for the DNA isolations with the Direct-zol RNA mini-prep kit (Zymo Research) according to the company's instructions and then subjected to a DNAse (Promega) treatment at 37 °C for one hour.

Generation of sequencing data

Sequencing of total-cellular DNA of *H. incana* 'Nijmegen' was performed by GenomeScan B.V., Leiden. A total of seven SMRT cells were used for sequencing on the Pacific Biosciences Sequel platform. Short read Illumina and 10X Genomics libraries with an insert size of approximately 500-700 bp were prepared with the NEBNext Ultra DNA Library Prep kit for Illumina and 10X Genomics ChromiumTM Genome v1 kit, respectively. These libraries were sequenced using the Illumina X10 platform (2 x 151 bp). RNA paired-end sequencing libraries with an average insert size of 254 bp were prepared using the Illumina TruSeq RNA sample prep kit with polyA mRNA selection and sequenced using the Illumina HiSeq 2500 platform (2 x 125 bp).

k-mer analysis

A histogram of k-mer frequencies of Illumina reads predicted to be of nuclear origin (see Methods S1) was generated using Jellyfish (v2.2.6) (Marçais & Kingsford, 2011), using a k-mer length of 21. The resulting histogram was provided as input to Genomescope (v1.0.0) (Vurture *et al*., 2017) to estimate genome size and heterozygosity.

Genomic assembly and annotation

The genome assembly and annotation process is more extensively described in Methods S1. In short, we generated an initial assembly based on the PacBio data only with Canu (Koren *et al*., 2017) and used it to bin the PacBio, 10X, and Illumina reads according to whether they originated from nuclear, organellar, or contaminant DNA. The bins were used to separately assemble the nuclear and organellar genomes, yielding a nuclear assembly consisting of hundreds of contigs and mitochondrial and chloroplast assemblies that were both represented by a single sequence. Nuclear contigs representing alternative haplotypes were removed using purge_dups (Guan *et al*., 2020), after which ARKS (Coombe *et al*., 2018) was used to scaffold the remaining contigs using the 10X data. Scaffolds were polished using Arrow (https://github. com/PacificBiosciences/gcpp) and Freebayes (Garrison & Marth, 2012), followed by a manual filtering step to obtain the final nuclear assembly.

Repeats in the assembly were masked using RepeatMasker (Smit *et al*., 2015) in combination with RepeatModeler2 (Flynn *et al*., 2020) before starting the annotation procedure. Nuclear genes were annotated by using EvidenceModeler (Haas *et al*., 2008) to generate consensus models of ab initio gene predictions, alignments of proteins from closely and distantly related plant species, and transcripts assembled from RNA-seq data. These models were manually filtered to obtain a final set of protein-coding genes.

Used datasets for comparative genome analyses

We mainly focused the comparative genome analyses on *H. incana*, *B. nigra*, and *B. rapa*, three species of the Brassiceae tribe of which all members underwent an ancient genome triplication (*Brassica rapa* Genome Sequencing Project Consortium *et al*., 2011; Lysak *et al*., 2005). For comparative gene analyses, we extended this group with the Brassicaceae species *Arabidopsis thaliana*, *Aethionema arabicum*, *Sisymbrium irio*, *Raphanus raphanistrum*, and *Raphanus sativus*. The latter two *Raphanus* species are also part of the Brassiceae tribe. Version numbers and locations of all genomes are listed in Table S14.

Analysis of pairwise gene synteny and long terminal repeat retrotransposons (LTR-RTs) in *H. incana*, *B. rapa*, and *B. nigra*

Analyses of pairwise gene synteny between scaffolds of *H. incana* and chromosomes of *B. rapa* and *B. nigra* were performed using the JCVI library (https://github.com/tanghaibao/jcvi) (v1.0.5) in Python. Orthologs were identified through all-vs-all alignment of genes with LAST (Kiełbasa *et al*., 2011), retaining reciprocal best hits only (C-score of at least 0.99). Hits were filtered for tandem duplicates (hits located within 10 genes from each other) and chained using the Python implementation of MCScan (Tang *et al*., 2008) to obtain collinear blocks containing at least four pairs of syntenic genes. Visualizations of collinearity between genomic assemblies were generated using custom scripts of JCVI.

Ks values of syntenic gene pairs were computed using the ks module of JCVI. Protein sequences of pairs were aligned against each other using MUSCLE (v3.8.1) (Edgar, 2004), after which PAL2NAL (v14) (Suyama *et al*., 2006) was used to convert protein alignments to nucleotide ones. Ks values for each pair were computed from the nucleotide alignments using the method of Yang and Nielsen, 2000 implemented in PAML (Yang, 2007) (v4.9). Times of divergence between species were estimated by dividing the median of the distributions of their Ks values by the rate of 8.22 × 10−9 synonymous substitutions per year that was established for Brassicaceae species based on extrapolation from the ancient triplication event in the Brassica clade (Beilstein *et al*., 2010).

Putative LTR-RTs were identified using LTRharvest (v1.6.1) (Ellinghaus *et al*., 2008) and LTR_finder (v1.1) (Z. Xu & Wang, 2007), after which LTR_retriever (v2.9.0) (Ou & Jiang, 2018) was run with default parameters to filter and combine the output of both tools into a high confidence set. LTR_retriever was also used to provide estimates of the insertion time of each LTR-RT. Parameters of LTRharvest and LTR_finder were set as recommended in the LTR_retriever documentation. Centromeric regions of the *B. nigra* assembly were obtained from Table S21 of the manuscript describing the assembly (Perumal *et al*., 2020).

Phylogenetic analysis of *H. incana*, *B. rapa*, and *B. nigra*

The longest isoforms of the nuclear genes of *H. incana*, *B. rapa*, *B. nigra*, and *A. thaliana* (outgroup) were provided to Orthofinder (version 2.3.11) (Emms & Kelly, 2019) to generate phylogenetic species trees. Orthofinder was run using the multiple sequence alignment (MSA) workflow with default parameters. The same analysis was performed using chloroplast genes. Trees were visualized using iTOL (version 6.3) (Letunic & Bork, 2021).

Comparative gene ontology analysis of eight Brassicaceae species

The longest isoforms of the genes of all eight Brassicaceae species described in the section "Used datasets for comparative genome analyses" were extracted using AGAT (version 0.2.3) (https://github.com/NBISweden/AGAT) and clustered into homology groups using the "group" function of Pantools version 2 (Sheikhizadeh Anari *et al*., 2018) with a relaxation parameter of 4. Groups were assigned GO slim terms of their associated *A. thaliana* genes (obtained from https://www.arabidopsis.org/download_files/GO_and_PO_Annotations/Gene_Ontology_Annotations/TAIR_GO_slim_categories.txt (last updated on 2020-07-01)) and GO terms assigned to protein domains of associated *H. incana*, *B. rapa*, and *B. nigra* genes using InterProScan (version 5.45-72.0) (Jones *et al*., 2014) (ran using the Pfam and Panther databases only). GO term enrichment tests were performed using the Fisher Exact test and the Benjamini-Hochberg method for multiple testing correction (Benjamini & Hochberg, 1995). *A. thaliana* genes were considered to be involved in photosynthesis, if they fulfilled one of the following conditions:

- Annotated with one of the following GO terms: "photosynthesis", "electron transporter, transferring electrons within the cyclic electron transport pathway of photosynthesis activity", or "electron transporter, transferring electrons within the noncyclic electron transport pathway of photosynthesis activity";
- Included in the KEGG pathways ath00195 (Photosynthesis), ath00710 (Carbon fixation in photosynthetic organisms), and ath00196 (Photosynthesis - Antenna Proteins);
- Protein products have been assigned the keyword "Photosynthesis" in the Swiss-Prot database;

The same criteria were used to retain photosynthesis-related genes of *H. incana* while filtering the gene annotation of the assembly (see Methods S1).

Investigating the mode of duplicated genes in *H. incana*

Dupgen_finder (Github commit hash 8001838) (Qiao *et al*., 2019) was run with default parameters to determine the mode of duplication for duplicated gene pairs in *H. incana*, using the genome of *A. thaliana* as an outgroup to detect pairs duplicated through whole-genome duplication. Pairs were allowed to be assigned to a single category only. Input files containing alignments of the protein sequences of *H. incana* aligned to themselves and those of *A. thaliana* were prepared using DIAMOND (version 0.9.14) (Buchfink *et al*., 2015).

Analysis of gene expression under high and low irradiance

Seeds of *H. incana* 'Nijmegen', *B. nigra* 'DG2', *B. rapa* 'R-o-18', and *A. thaliana* Col-0 were germinated on a peat-based potting mixture for nine days under an irradiance of 200 μmol · m^{-2} · s^{-1}. Twelve seedlings per species were then transferred to 2L pots filled with a peat-based potting mixture enriched with perlite and 2.5 g/L Osmocote ®Exact Standard 5-6M slow-release fertiliser (ICL Specialty Fertilizers, Geldermalsen, The Netherlands).

Plants were germinated and grown in a climate chamber with a photoperiod of 12 hours and day and night temperatures of 23 and 20 °C, respectively. Humidity and CO2 levels were set at 70% and 400 ppm. The chamber was equipped with high-output LED light modules (VYPR2p, Fluence by OSRAM, Austin, USA). Six plants per species were assigned to a high light (HL) treatment of 1800 $\mu mol \cdot m^{-2} \cdot s^{-1}$ and the remaining six to a low light (LL) treatment of 200 $\mu mol \cdot m^{-2} \cdot s^{-1}$. Irradiance uniformity was very high for both HL and LL treatments, with a U2 value of 0.93. Plant positions were randomised across growing areas. Plants were watered with the same custom nutrient solution as the one used in the measurements of photosynthesis rates, daily for the LL treatment and twice a day for the HL treatment.

Twenty-eight days after sowing, one young fully adapted leaf from each plant was selected, excised, and snap-frozen in liquid nitrogen. Leaf samples were crushed with a mortar and pestle cooled with liquid nitrogen and further homogenised with glass beads for 2min at 30Hz in a MM300 Mixer Mill (Retsch GmbH, Haan, Germany). Total RNA was extracted with the RNeasy Plant Mini Kit (QIAGEN N.V., Venlo, The Netherlands) according to manufacturer's instructions and then subjected to a RQ1 DNAse treatment (Promega Corporation, Madison, U.S.) at 37 °C for 30 minutes. We validated the total removal of DNA by means of a no-reverse transcriptase PCR reaction on all RNA samples. The RNA quality was assessed for purity (A260/A280) with a Nanodrop 2000 spectrophotometer (Thermo Fisher Scientific Inc., Waltham, U.S.) and for possible RNA degradation by means of a visual inspection of the RNA on a 1% agarose gel. cDNA was then synthesized from 2 µg total RNA (measured by spectrophotometer) with the SensiFAST™ cDNA Synthesis Kit (Meridian Bioscience, Cincinnati, U.S.) according to manufacturer's instructions.

To examine the expression of both single-copy and multi-copy photosynthesis and/or photoprotection-related genes (Table S15), species-specific RT-qPCR primers were designed with the following criteria: the PCR fragment size had to range between 80 and 120 bp, the maximum difference in melting temperature between primers of the same pair had to be 0.5 °C, and overall melting temperatures had to be comprised between 58 and 62 °C. Primers were designed to target a region of the gene as similar as possible in all species. Additionally, for multi-copy genes, the primer pair had to bind to all copies of a particular gene in one species. RT-qPCR reactions were performed with SYBR green on a CFX96 Real-Time PCR Detection System (Bio-Rad Laboratories Inc., Hercules, U.S.). The efficiency of each designed primer set was assessed by means of a standard curve, and only primer sets with efficiencies ranging between 90% and 110% were used. All primer sequences can be found in Table S16.

Gene expression was normalized to the reference genes *ACT2*, *PGK*, *UBQ7* and *APR* (Joseph *et al*., 2018; Løvdal & Lillo, 2009) using the delta-Ct (dCt) method (Livak & Schmittgen, 2001). Normalized gene expression values were calculated as 2^{-dCt}. For the statistical analysis, we performed two-way ANOVA on the dCt values with the copy number and light treatment as grouping variables for the multi-copy genes and species, and light treatment as grouping variables for the single copy genes. A Kenward-Roger approximation for the degrees of freedom was used and a post-hoc test was subsequently performed with Tukey's range test, with the significance threshold set at (α = 0.05).

Availability of data and material

Raw sequencing data of *H. incana* can be found on the repository of the National Center for Biotechnology Information (NCBI) (BioProject ID: PRJNA612790). The genome assembly of *H. incana* has been deposited at DDBJ/ENA/GenBank under the accession JABCMI000000000. The version described in this manuscript is version JABCMI010000000. The genome assembly and annotation files, as well as a Jbrowse instance of the genome, can be accessed at https://www.bioinformatics.nl/hirschfeldia.

Supplementary Figures

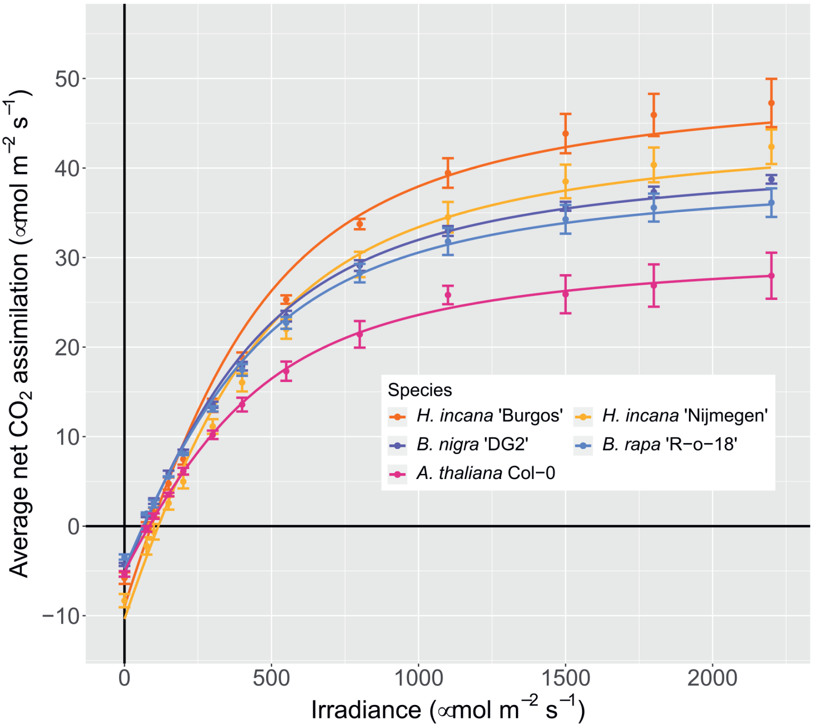


Figure S1: **Net CO_2 assimilation rates as measured with the LiCor 6400XT.** Light-response curves for *Hirschfeldia incana*, *Brassica rapa*, *Brassica nigra*, and *Arabidopsis thaliana* accessions adapted to high levels of irradiance. Each point represents the mean net CO_2 assimilation value of three (*B. rapa*) or four leaves coming from independent plants. The lines represent non-rectangular hyperbolas fitted on the raw data to estimate daytime dark respiration (Rd). Error bars represent the standard error of the means.

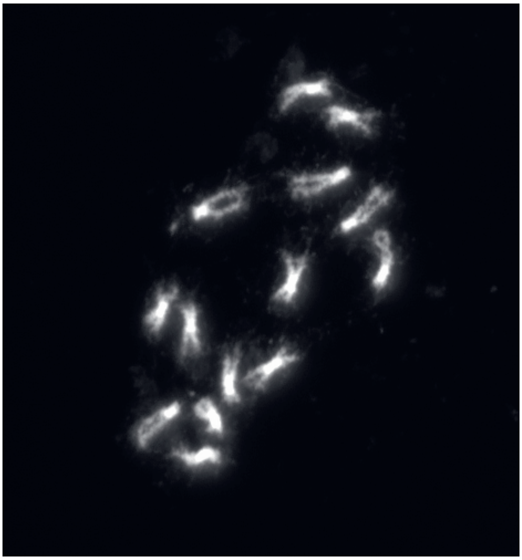


Figure S2: **Root tip smears of *Hirschfeldia incana*.** An example of 4′,6-diamidino-2-phenylindole (DAPI)-stained chromosomes from one metaphase mitotic cell. Fourteen chromosomes can be observed, meaning that *H. incana* has a haploid chromosome number of n = 7.

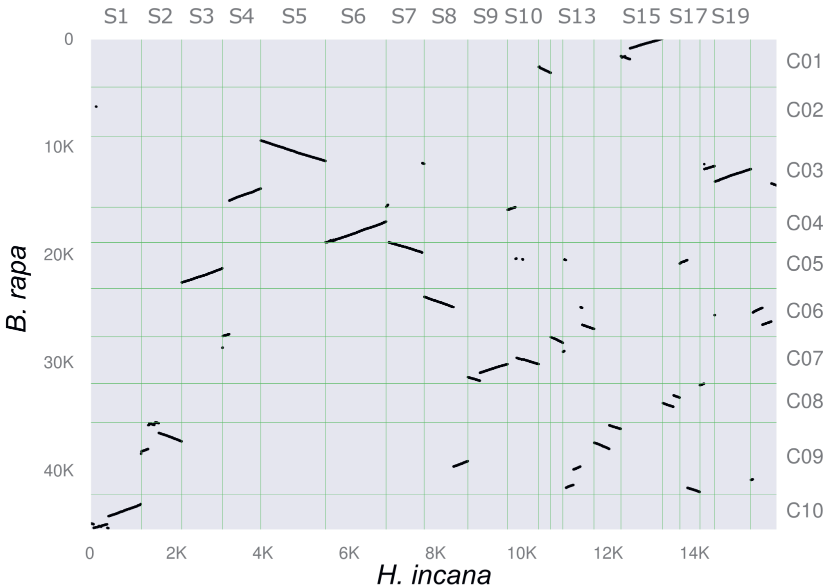


Figure S3: **Orthologous syntenic blocks between the genomes of *Hirschfeldia incana* and *Brassica rapa*.** Dots indicate pairs of syntenic orthologous that are found in the same order in both genomes according to sequence position. Only the twenty largest scaffolds of *H. incana* (43.6% of the assembly) are shown for clarity. Axes labels correspond to the total number of genes annotated on the sequences (left and bottom) and identifiers of the scaffolds (top) or chromosomes (right).

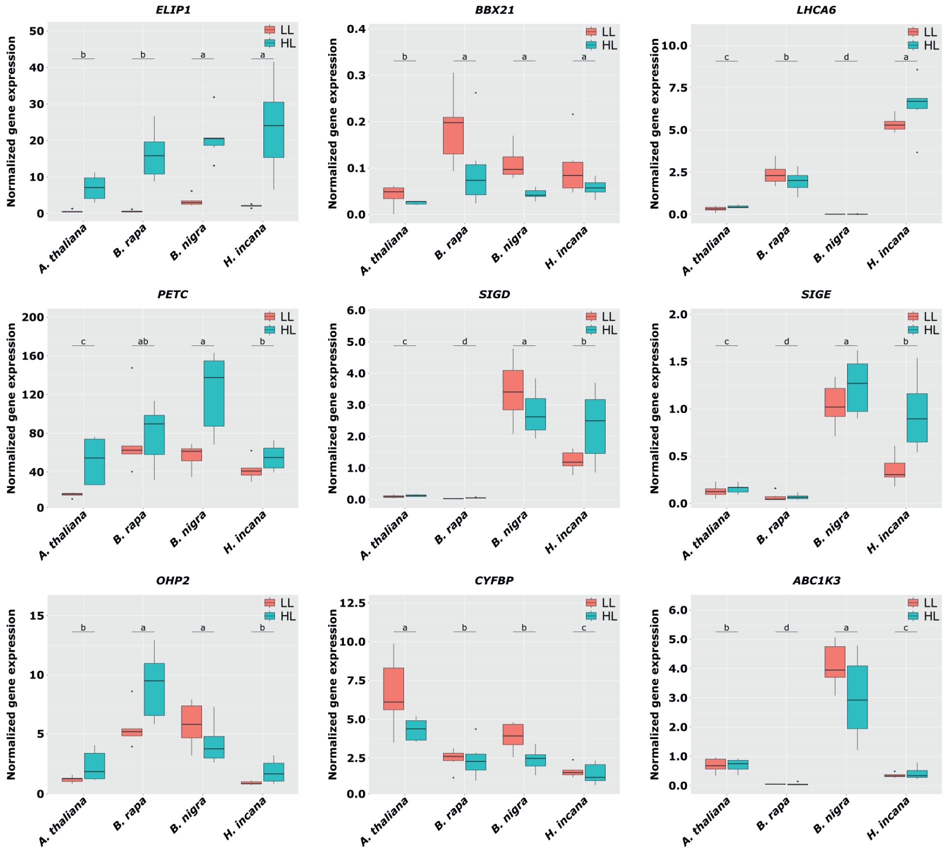


Figure S4: **Normalized relative gene expression per species and light treatment for the set of multi-copy photosynthesis-related genes.** Boxplots depict normalized gene expression levels of *A. thaliana*, *B. rapa*, *B. nigra* and *H. incana* grown in 200 µmol m^{-2} · s^{-1} (LL) and 1800 µmol m^{-2} · s^{-1} (HL) for genes for which there is variation in gene copy number between these species. Graph titles indicate gene names based on the *A. thaliana* gene nomenclature. Letters indicate statistical differences between species based on a two-way ANOVA and Tukey posthoc test, testing for light treatment and species as variables.

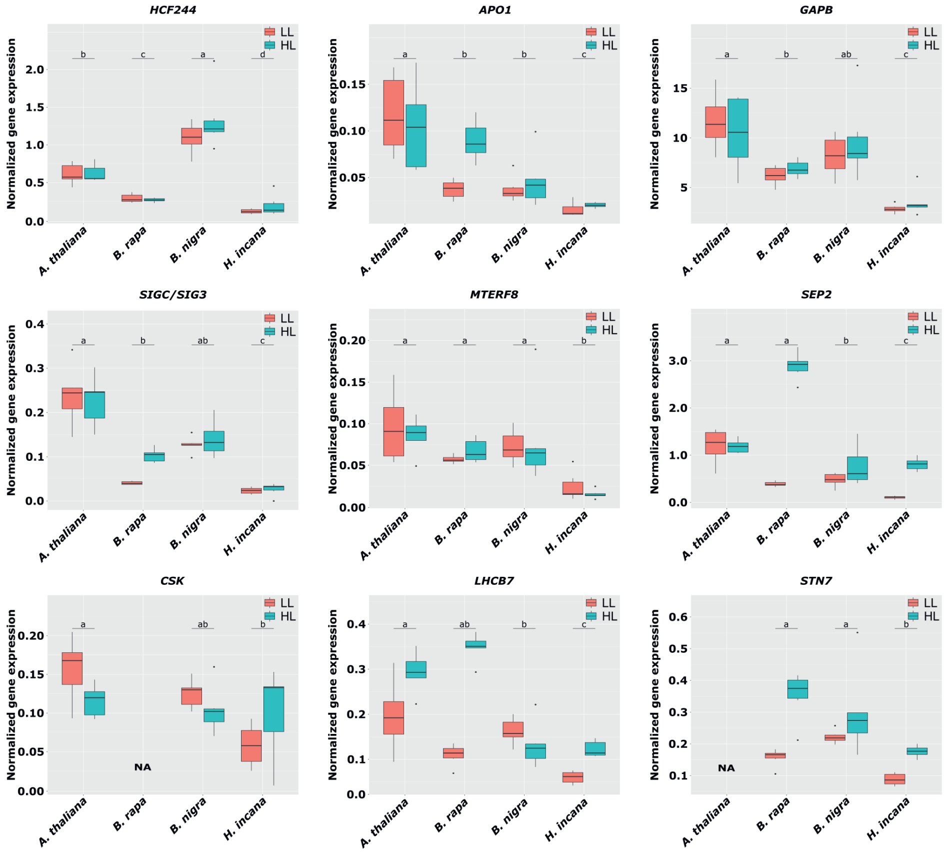


Figure S5: **Normalized relative gene expression per species and light treatment for the set of single-copy photosynthesis-related genes.** Boxplots depict normalized gene expression levels of *A. thaliana*, *B. rapa*, *B. nigra* and *H. incana* grown in 200 μmol m^{-2} · s^{-1} (LL) and 1800 μmol m^{-2} · s^{-1} (HL) for genes for which there is variation in gene copy number between these species. Graph titles indicate gene names based on the *A. thaliana* gene nomenclature. Letters indicate statistical differences between species based on a two-way ANOVA and Tukey posthoc test, testing for light treatment and species as variables.

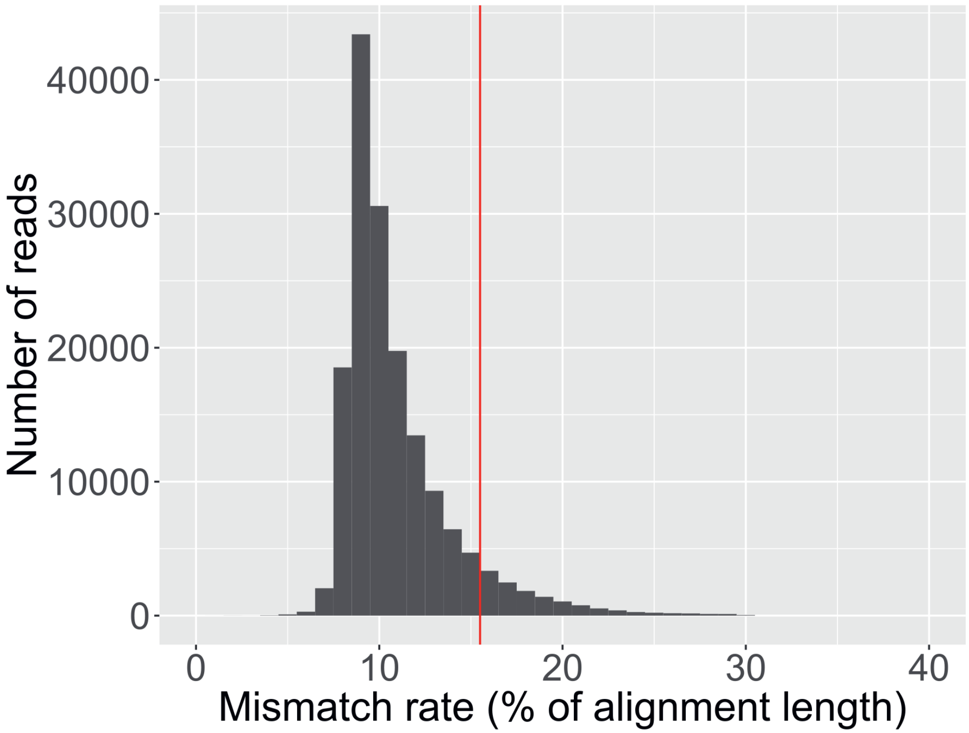


Figure S6: **Histogram of mismatch rates of PacBio reads aligned against the chloroplast and mitochondrial assembly of *Hirschfeldia incana*.** The cut-off of 15%, used to distinguish organelle-derived reads from reads originating from nuclear regions that contain organellar inserts, is depicted by the red vertical line.

Chapter 6

General discussion

Introduction

In this thesis I have investigated genetic aspects of plant adaptation to their environment. At the start of the work presented in this thesis, the project of which this PhD program was part of, was entitled 'Exploiting copy number variation for rapid improvement of abiotic stress tolerance in crops'. The main premise of the project was that copy number variation (CNV) appears to be a highly dynamic type of genetic variation with a large, and underappreciated, role in regards to the evolution of plant adaptation to its environment. This was further explored in **Chapter 2**. Experimental evidence from different organisms demonstrated common causes and mechanisms underlying de novo CNV formation that are more prone to errors in stressful environments. Similarly, experimental evidence in plants (e.g. (DeBolt, 2010)) further strengthened the idea that de novo CNV can quickly arise in stressful environments in plants. Meanwhile, next generation sequencing (NGS) technology became more accessible, which provided the opportunity to further examine the role of CNVs in relation to plant adaptation. By incorporating CNVs alongside the more commonly used genetic variants such as SNPs and InDels, I attempted to test their contribution in an as unbiased manner as possible.

Despite the initial specific focus on CNV, that initiated most of the rationale behind the approaches of the experimental chapters, the topic of CNV did not come out as prominently in the experimental chapters as initially anticipated. The focus of the thesis, therefore was shifted more towards the impact of mutation and natural genetic variation in general on plant adaptation to their environment. Nevertheless, in this General Discussion, I will first address the role of CNV to plant adaptation and proceed with further insights on the rapid adaptation to severe zinc stress. Then I shift the topic to study of plant adaptive traits using natural variation with a strong focus on the use of genome-wide association studies. Finally, I discuss the potential for better understanding the high-photosynthetic trait of *Hirschfeldia incana*, a close relative to several important crop species, as a possible avenue to improve photosynthesis.

The role of CNV to rapid adaptation

One hypothesis of this thesis was that spontaneous CNV formation could contribute to rapid adaptation to the environment. As discussed in the General Introduction (**Chapter 1**) and **Chapter 2**, plant stress may induce CNV formation through the activation of TE activity, DNA replication and repair processes, and recombination events. Additionally, studies in human cell lines and yeast have shown that CNV is more likely to occur in highly transcribed regions (Hull *et al*., 2017; Lauer & Gresham, 2019). In these regions, collisions between the transcription and replication machinery are known to be particularly mutagenic and can lead to the formation of CNV hotspots (Aguilera & García-Muse, 2013; Hull *et al*., 2017; Mansisidor *et al*., 2018; Sankar *et al*., 2016; Wilson *et al*., 2015). The potential link between transcriptional activity and CNV formation suggests that stress may promote CNV formation at loci with strongly induced expression in response to environmental stressors, thereby accelerating the rate at which mutations are generated, including adaptive variants (Hull *et al*., 2017). However, the connection between transcriptional activity and CNV formation requires additional genomic features or a loss of control when a critical stress level is achieved, because otherwise CNV would also be expected to occur more frequently for highly transcribed house-keeping genes, which is not observed. For instance, in yeast, CNV induction of a transgene that provides increased nickel tolerance is also dependent on H3K56 acetylation of chromatin (Hull *et al*., 2017; Whale *et al*., 2022). Currently, it is not known whether high transcriptionally active loci are also more prone to CNV in plants, but this phenomenon has been observed in both prokaryotes and eukaryotes, suggesting that it could represent a relevant mechanism promoting adaptive mutations in plants.

One of the reasons to use zinc-stress in the experimental evolution approach was that CNV for metal homeostasis-genes is frequently observed in species that have adapted high metal concentrations. For instance, adaptation to high concentrations of zinc has evolved in the closely-related species *Arabidopsis halleri* and *Noccaea caerulescens* (Krämer, 2010). These species are both hypertolerant to zinc, and are also capable of hyperaccumulating it. The evolution of these traits involved copy-number expansions of several metal homeostasis genes (Dräger *et al*., 2004; Hanikenne *et al*., 2008; Krämer, 2010; J.-S. Peng *et al*., 2021). For instance, duplication of *HEAVY METAL ATPASE 4* (*HMA4*) is involved in cadmium and zinc hypertolerance in *A. halleri* and *N. caerulescens* (Craciun *et al*., 2012; Ó Lochlainn *et al*., 2011; Suryawanshi *et al*., 2016). There is further CNV among different populations in *N. caerulescens*, for instance four copies of *HMA4* are present in the hyperaccumulator populations populations Ganges and St-Félix-de-Pallières, three copies in the intermediate population Prayon and two copies in the lowest hyperaccumulating population Puente Basadre (Craciun *et al*., 2012). This pattern where genes involved in metal tolerance and hyperaccumulation have been duplicated in a species, but are not fixed in their numbers across the species, is more commonly observed. In *A. halleri* for instance, duplication of *ZRT-IRT-like PROTEIN 6* (*ZIP6*) is involved in fine-tuning of metal-homeostasis, and there is CNV between different ecotypes (Spielmann *et al*., 2020). Also for *N. caerulescens*, I have determined the copy number of ZIP6 and detected CNV, ranging from 1 to 3 copies among different ecotypes (unpublished work). These variable numbers among different ecotypes likely reflect differences in genomic dynamics during local adaptation.

The contribution of CNV to rapid adaptation to the environment was experimentally addressed in **Chapter 3**. No evidence or indication of increased CNV formation was observed in plants exposed to severe salinity or zinc stress. In fact, only a very small number of spontaneous CNVs were detected (five de novo CNVs out of a total of 200 plants). Therefore, based on my own experimental data, I cannot conclude that CNV rates are increased in *A. thaliana* upon stress exposure. However, there are two important things to take into consideration in regards to this conclusion. Firstly, the experimental set-up utilized in **Chapter 3** does not allow the formal testing of the spontaneous mutation rate due to the strong selection imposed by the experimental set-up. If CNVs are more likely to be deleterious, strong purifying selection against them could reduce the number of variants that are accumulated over five experimental generations. In such case, the observed number of accumulated CNVs does not accurately reflect the actual number of mutations that occurred. Therefore, for examining the spontaneous mutation rate, a single-seed mutation accumulation (MA) approach would be more suitable. This is also observed in an experiment using *Daphnia pulex*, in which a MA experiment was conducted in the presence of copper and nickel, and an experiment in which a population was propagated under competitive conditions (Chain *et al*., 2019). The CNV mutation rate was about four-fold higher in response to copper and nickel compared to the control, but was reduced to only 0.4x the rate of control under competitive conditions (Chain *et al*., 2019). Nevertheless, in the control treatments of the experiment in **Chapter 3**, the selection pressure against CNVs is likely lower than under severe stress, but even there hardly any CNVs are observed.

A second, and perhaps even more significant, confounding factor in this research is the difficulty in accurately detecting CNVs, especially gains of copies, using short-read Illumina sequencing data, despite the use of the bioinformatics tool Hecaton (Wijfjes *et al*., 2019). This challenge is, in part, due to the reliance on a single reference genome in the alignment process. Alignment tools attempt to match each sequencing read to a specific location on the reference genome. However, when genomic sequences are absent in the reference genome, these reads are often discarded from further analysis. While this removal of non-aligning reads is beneficial in eliminating contaminant reads (from bacteria, fungi, etc.) present in the original sample, it hinders the detection of the majority of insertions. These difficulties in detecting certain types of CNVs using short-read sequencing technology are also encountered in this study. For example, when Hecaton was applied to the DartMap sequencing data (**Chapter 4**), we observed that significantly more deletions were detected than insertions and duplications (Figure 1). This outcome is unexpected, as CNVs are called against the Col-0 reference genome, and there is no reason to assume that Col-0 contains more genomic sequences than other *A. thaliana* plants. Gains of copies, which are of particular interest as potential adaptive mutations, are currently not accurately detected in the analysis, which makes it challenging to assess their contribution to adaptation.

Nowadays, the challenges of accurately detecting CNVs are mostly overcome by the use of long-read sequencing technologies, which offer two main advantages over short-read sequencing technology for CNV detection. Firstly, the increased sequencing read length in long-read sequencing makes it much easier to detect CNVs, especially in repetitive sequences, as individual reads can span over (parts of) duplicated regions. This ability enhances the accuracy of CNV detection in complex genomic regions. Secondly, long-read sequencing is particularly suited for generating de novo genome assemblies, which can be valuable for detecting CNVs and other structural variants (SVs) through comparisons

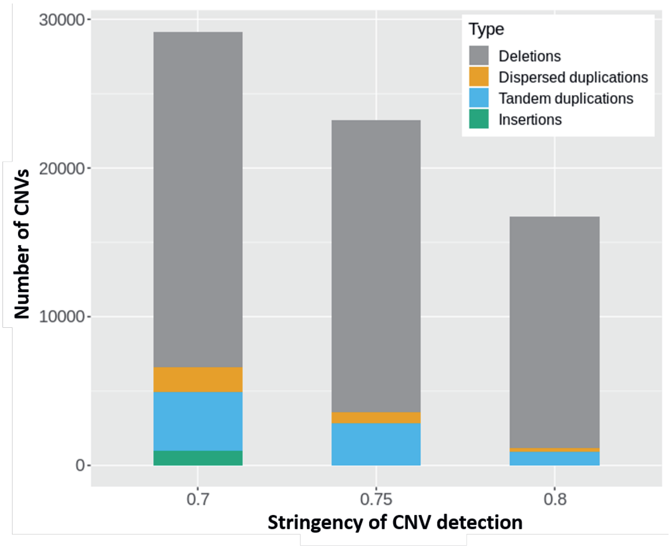


Figure 1: **Different types of CNVs detected in the 192 DartMap accessions.** The stringency of CNV detection by affects how many CNVs are found. Additionally, the tool is better suited to detect deletions (grey) than dispersed and tandem duplications (orange/blue respectively) or insertions (green). Figure S11 from Chapter 4.

of such assemblies. Although long-read sequencing still faces challenges in terms of relatively lower throughput and higher costs compared to short-read sequencing, its implementation in studies similar to those conducted in **Chapter 3** and **Chapter 4** is becoming increasingly accessible. For instance, in a recent study, 72 de novo genome assemblies of genetically diverse *A. thaliana* accessions were successfully constructed using long-read sequencing, revealing that only 52% of genes are present in all accessions (Lian *et al*., 2023). It is unlikely that long-read sequencing will completely replace short-read sequencing, but their integration will significantly advance the detection of CNVs and SVs in diversity panels. For instance, CNV detection tools based on short-read sequencing data are well equipped to detect deletions. Therefore, using a pan-genome derived from long-read sequencing, rather than a single reference genome, could be advantageous for the detection of CNVs in natural variation panels that have been sequenced using short-read sequencing. Long-read sequencing is especially useful for the detection of rare variants, such as those that may have arisen during the experimental evolution approach. Therefore, using long-read sequencing on a selection of plants from the zinc experiments could provide further valuable insights in the contribution of CNVs to short-term stress adaptation.

The rapid adaptation of increased tolerance to excess zinc

Increased mutation rates in response to various stresses have been observed in various organisms (Belfield *et al*., 2021; Hull *et al*., 2017; C. Jiang *et al*., 2014; Z. Lu *et al*., 2021; Matsuba *et al*., 2013; Sniegowski *et al*., 1997). However, the adaptive value related to an increased mutation rates has not been studied much in plants. While MA experiments are more suitable for estimating the effect of stress on mutation rates, an experimental evolution approach allows for a direct link between spontaneous variation and adaptive phenotypes. Prior to starting the experiment, my hypothesis was that the likelihood of generating sufficient spontaneous genetic variation leading to substantial adaptation within only five generations was low. Nevertheless, rapid and significant increases in zinc tolerance were observed in one of the experimental populations (ZSI2). In addition to this clear example of zinc tolerance that evolved during the experiment, several individual plants from other replicate populations also exhibited increased levels of zinc tolerance. Although these cases have not been studied further, they suggest the independent development of increased zinc tolerance in replicate populations. Studying such other examples in independent replicate populations is important to further establish if the observed example of rapid adaptation was ‘a chance occurrence’, or if adaptation to severe zinc stress indeed occurred multiple times independently.

The rate of adaptive phenotype generation depends on various factors, including the genetic complexity of the trait under investigation and the rate of environmental change. Typically, adaptation to a strong and sudden environmental change via mutation occurs through fixation of a limited number of relatively large-effect mutations that typically arrive early on (Barrick & Lenski, 2013). Adaptation to more gradual environmental changes may instead involve a higher number of mutations with relatively small effect that occur at less predictable times, but these can result in a higher final fitness (Collins & De Meaux, 2009). The experimental evolution approach used here for the zinc stress treatments reflects best the dynamics of a sudden and strong environmental change, thus the selection of a single large-effect size mutation is in line with what could be expected under such environment. Typically, such large-effect mutations have trade-offs in other environments. This is also the case for the zinc-tolerant plants. Under control conditions ZSI2 plants grow slightly slower compared to ZC-1 adapted plants, while under high salinity they perform considerably worse.

Studying the genetic basis of plant traits using genome wide association analysis

Understanding the genetic basis of plant adaptive traits is fundamental to devising strategies that can optimise crops. Natural variation panels provide an excellent tool to study plant responses to their environment and help dissect complex genetic traits (Tibbs Cortes *et al*., 2021). The use of natural variation panels relies, in part, on the notion that plants have locally adapted to their environment by accumulating beneficial alleles. GWAS is being routinely applied in many different plant species, including *A. thaliana* (Korte & Farlow, 2013; Liu & Yan, 2019). This is partly due to the availability of abundant genotyping data and the relative ease of conducting the analysis. For instance, web applications such as GWAPP (Seren *et al*., 2012) have made GWAS widely accessible to the *A. thaliana* community. Despite the relative simplicity of running such an analysis, only a small subset of studies has actually resolved the causal gene(s) underlying their associations (Sasaki *et al*., 2021). Not every study aims to achieve this outcome, but in my impression, this is also partly because there are several experimental (and conceptual) considerations that may vary on a case-by-case basis and are open to interpretation. It is important to be aware of such considerations since time constraints often only allow for the study of a few associations, as gene validation tends to be a tedious and time-consuming process. It is beyond the scope of this discussion to provide a complete and comprehensive overview of all the considerations and factors that could, or perhaps should, be taken into account when conducting a GWAS, but several important ones will be highlighted.

The choice for population

The decision for what natural variation panel to use for GWAS will impact the outcome. Most early GWAS in *A. thaliana* made use of globally collected natural variation panels, such as the RegMap and HapMap populations (Horton *et al*., 2012; Y. Li *et al*., 2010). The choice of these populations was based on sampling plants across a diverse range of environments and the premise that including geographically distant accessions would maximise the genetic variance within the population (Korte & Farlow, 2013). However, a trade-off of higher genetic variance is a lower proportion of variants shared by multiple accessions, and therefore allelic heterogeneity is more likely to occur. Allelic heterogeneity means that the same phenotype can be caused by different mutations within the same gene. It may prevent the identification of causal loci through GWAS because the statistical approaches used for GWAS typically calculate variant-phenotype association scores.

If causal variants do not occur at a sufficiently high frequency within the population, they will be filtered out by using a minor allele frequency cut-off correction or may lack sufficient statistical power to reach significance. A local population with a lower degree of genetic variance, and therefore allelic heterogeneity, would thus be less affected by this issue. For instance, when considering flowering time, a trait that has proven quite difficult to map in global populations (Atwell *et al*., 2010), the DartMap demonstrates a strong association because a single *fri* knockout allele causing early flowering occurs at a relatively high frequency. In the last decade, several more local natural variation panels have become available, such as those from Sweden (Long *et al*., 2013), the Iberian Peninsula (Tabas-Madrid *et al*., 2018), and a small region in France (Frachon *et al*., 2018). Similar to what we observed with the DartMap panel (**Chapter 4**), these panels exhibit relatively high genetic diversity despite being located in relatively small geographic areas. This suggests that a large proportion of the genetic variation found in *A. thaliana* is likely (close to) neutral in most environments, and its distribution influenced mainly by genetic drift. The proportion of locally adaptive alleles may only constitute a much smaller fraction of the total genetic variation.

Adaptive or cryptic variation?

For trait discovery approaches such as GWAS, it is not relevant whether genetic variation is adaptive or neutral in its native environment, as long as it translates to phenotypic variation for the studied traits. For example, we discovered that allelic variation of *FSD3* underlies a differential response to iron deficiency. The distribution of *FSD3* alleles on Dutch soil maps is not correlated to the soil type or soil parameters on which accessions are collected. Although the general soil maps of the Netherlands may inaccurately represent the actual soils on which the accessions were collected (e.g. between pavements and in people's gardens), there are no clear indications to assume that allelic variation at *FSD3* plays an important role in local adaptation. Instead, it may represent a case of standing genetic variation, with the phenotypic implications only becoming apparent in a specific environment we provided experimentally. A GWA approach can even be entirely unsuitable in *A. thaliana* for loci that are under strong positive selection. In **Chapter 4**, one specific genotype is observed with a loss-of-function *ga5* allele that has spread across a considerable distance (>100 km) and is recurrently found at a relatively high frequency within the Netherlands. The loss-of-function *ga5* allele results in dwarfism, which appears to confer a benefit in windy conditions. This aligns with the windier conditions typically found in the general area where this genotype is predominantly observed, suggesting a potential example of local adaptation. However, the same pattern of genotype distribution could have emerged from genetic drift. If, for instance, the loss-of-function *ga5* allele is nearly neutral in terms of fitness, the absence of selection against this allele could also contribute to its spread. Regardless of whether this specific case represents an example of local adaptation, the same pattern of observing genetically nearly identical plants within a localized area is expected when a new spontaneous mutation provides a strong selective advantage. In such cases, both the trait and the causal mutation are entirely linked to the genetic background in which it arose, making it impossible to discriminate the causal mutation from all other genetic variants that were already present before the occurrence of the new mutation (Figure 2).

Outliers and unbalanced alleles increase the risk of false positive associations

An important consideration is whether the phenotypic data should be cleared of outliers or not. Phenotypic outliers can represent accurate and realistic phenotypes for specific plants. It may thus seem counterintuitive to exclude such outliers from the analysis when there is no reason to question the quality of the measurements. However, outlier data can significantly affect the outcome of a GWAS and lead to many false positive associations (Alvarez Prado *et al*., 2019). Loci with highly unbalanced allele frequencies are particularly susceptible to false positive associations resulting from outliers (Alvarez Prado *et al*., 2019). While removing outlier data carries the risk of potentially missing rare alleles that could have a strong impact on the phenotype, variants with unbalanced frequency distributions are still more likely to yield inflated test statistics, even after outliers are removed (Shen & Carlborg, 2013).

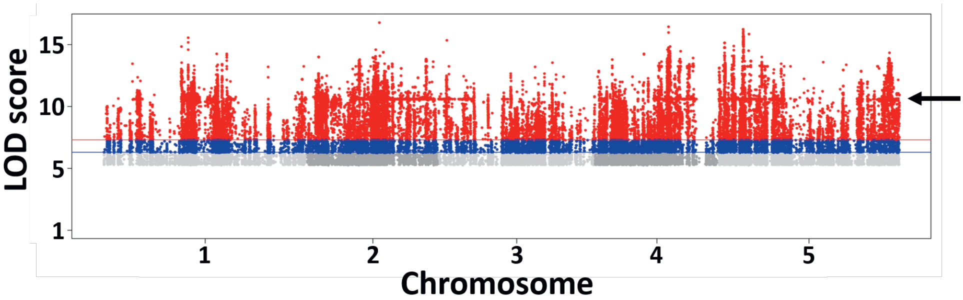


Figure 2: **Genome wide association (GWA) analysis on stem length in the DartMap without kinship correction.** Here, a loss-of-function mutation in GA5 results in dwarfism. The mutation arose in a particular genotype and subsequently spread across coastal areas of the Netherlands. Because the mutation is genetically linked to many other variants all across the genome, it cannot be discriminated from them on a statistical basis. In a regular GWAS, a kinship (relatedness) correction is used to prevent against these type of associations. Here, an example is provided of such an analysis without using a kinship correction. The association scores are heavily inflated and a horizontal line, indicated with a black arrow, is formed on parts of the chromosomes around approximately LOD = 11, which represent variants that are exclusively present in the accessions that share the same *ga5* mutation, including the *ga5* mutation itself.

To investigate whether this phenomenon occurs in the DartMap panel when using the software GEMMA, I randomly simulated normally distributed phenotypic data and performed 1000 permutations of these random data for GWA analyses. For each permutation, I identified the variant with the highest association score to examine whether alleles with relatively low MAF were overrepresented among the top associations compared to random chance (Figure 3). In this simulation, 62.8% of the 1000 highest associated variants had a MAF ≤ 10%, while within the DartMap panel, only 29.9% of variants have a MAF ≤ 10%. This example demonstrates that variants with a MAF ≤ 10% are about twice as often detected as false positives as expected if the allele frequency had no effect on the association score. The presence of outlier data would likely exacerbate this problem. Therefore, allele frequencies should be carefully considered when deciding which associations to pursue further. While they may present causal associations, investigating them further may be a risky endeavour.

The threshold of significance

Similarly, a common consideration is the choice of the threshold of significance when determining which associations to pursue further. When testing many associations (often in the millions) between a phenotype of interest and individual variants, a stringent multiple-testing correction must be applied to avoid false positives. Most commonly these significance thresholds are determined either by limiting the false discovery rate (FDR)

(Benjamini & Hochberg, 1995), by using a Bonferroni correction that divides the desired significance level by the total number of independent tests, or by employing permutation tests. Both FDR and Bonferroni corrections can be overly conservative, leading to false negatives (Gupta *et al*., 2019). Permutation tests are considered the gold standard for setting threshold levels (Alvarez Prado *et al*., 2019), but they can be computationally intensive, especially when performed for each individual phenotype. It is essential to keep in mind that the association analysis is only capable of highlighting genomic regions that correlate with the phenotype of interest. The permutation analysis (Figure 3) demonstrates that the allele frequency affects the likelihood of inflated associations. Consequently, loci with higher MAFs may require less stringent threshold levels compared to loci with relatively lower MAFs. It is important to note that an association analysis alone cannot establish causality, regardless of how statistically significant the association may be. Therefore, a GWAS should be regarded primarily as an exploratory analysis, and the significance threshold should be viewed as a tool that aids in prioritizing genomic loci that appear most promising for further validation.

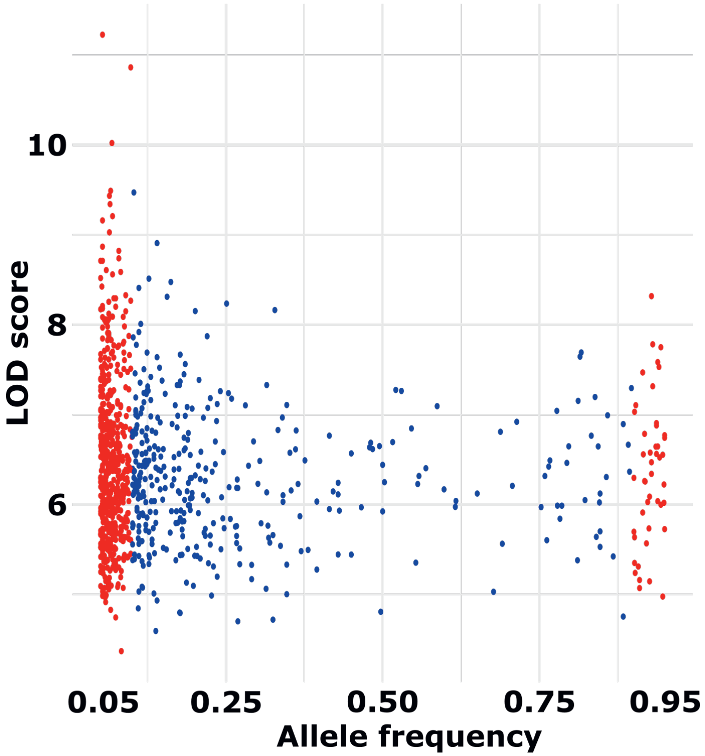


Figure 3. **Unbalanced allele frequencies are more prone to inflated test statistics.** Here, the highest associated variants (variants with MAF ≤ 10% in red, MAF ≥ 10% in blue) from 1000 permutations of randomly simulated data are plotted based on their allele frequency.

Detection of some signals of false positive hits

Careful examination of associated variants and their surrounding genomic regions can aid in detecting false positive signals. For instance, single variants with strong association scores but no other nearby variants forming a discernible peak should be treated with caution. Genetic variants are expected to exhibit some degree of linkage disequilibrium (LD) with other nearby variants. Consequently, these genetically linked variants are correlated with each other, and their statistical association should also display correlation, resulting in characteristic peaks in a GWAS rather than “singletons”. Such singletons are, in *A. thaliana*, quite frequently associated with alleles that have a relatively high occurrence of heterozygosity within the population. It is common for accessions to be either homozygous for the reference allele or labelled as ‘heterozygous’ despite nearby variants in those accessions being fully homozygous. These ‘heterozygous’ variants often arise from sequencing reads originating from genomic regions absent in the Col-0 reference genome but mapping to similar, non-allelic sequences in the reference genome, thus creating a false signal of heterozygosity. CNV of for instance transposable elements, variable tandem repeats, and other types of structural variation can be sources of such artifacts. These artifacts can often be identified by comparing read alignment patterns among samples with contrasting alleles. When associations are observed with such alleles, they may indicate cases where the presence or absence of the duplicated region associates with a trait of interest, as illustrated in Figure 4. It is very challenging though, to pinpoint the precise genomic region of the duplicated sequence, and thus identify the allelic variation causing the association.

Gene validation is crucial to establish causality

Gene validation experiments are essential for establishing causality of candidate genes. However, relatively few plant GWA studies incorporate gene validation experiments into their experimental design (H. Liu & Yan, 2019). Initial gene validation experiments may involve gene expression analyses of accessions with contrasting alleles and phenotypic evaluations of candidate gene knock-out, knock-down or overexpression lines (Alseekh *et al*., 2021; H. Liu & Yan, 2019). While these experiments can prioritize candidate genes, they cannot establish causality on their own. In *A. thaliana*, knock-out mutants are commonly used for gene validation experiments. If a knock-out mutant of a candidate gene exhibits a phenotype related to the trait of interest, it can be considered an indication of the gene’s involvement in the trait. However, it does not demonstrate the presence and impact of natural allelic variation for that specific gene on the trait.

Especially for highly complex polygenic traits, such as plant growth or photosynthesis, observing a phenotype in a knock-out mutant may not provide the most informative evidence since knock-out alleles of many genes will likely have (indirect) effects on such traits. Similarly, the absence of a phenotype in a knock-out mutant does not automatically imply that the gene is not involved. Until recently, T-DNA insertion lines were commonly used for such analysis due to their easy availability. However, a

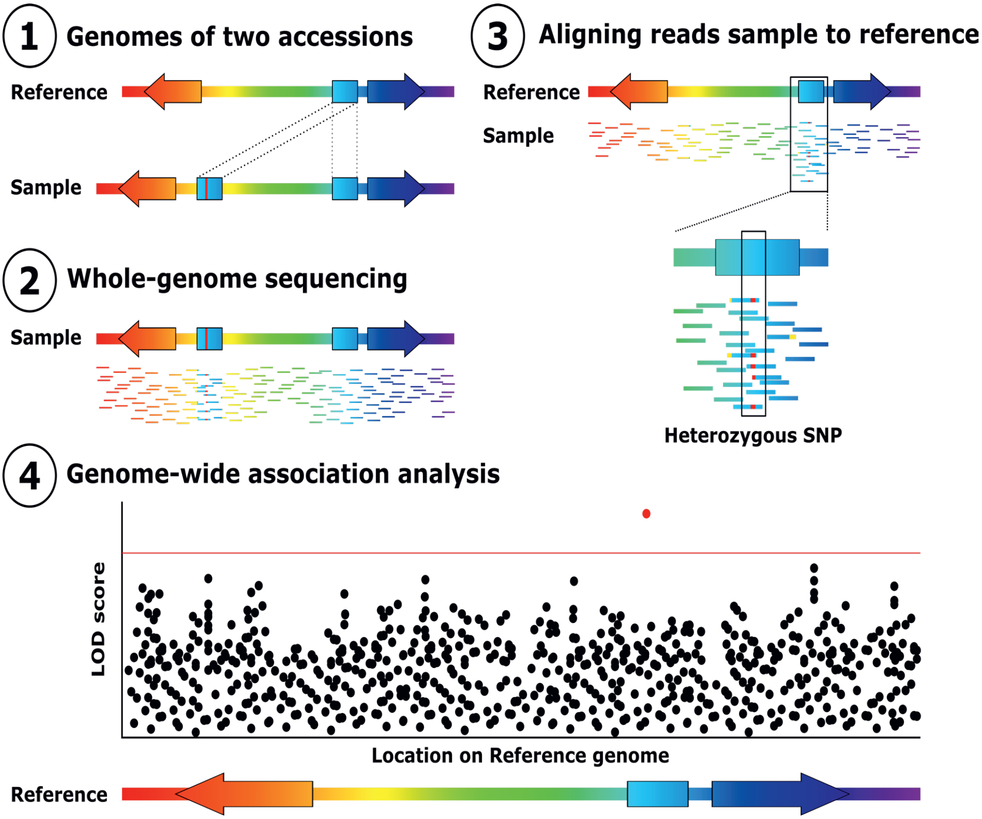

1
Genomes of two accessions
Reference
Sample
2
Whole-genome sequencing
Sample
3
Aligning reads sample to reference
Reference
Sample
Heterozygous SNP
4
Genome-wide association analysis
LOD score
Location on Reference genome
Reference

Figure 4: **Duplications can give rise to false signals of heterozygosity that can result in singleton associations in genome-wide association analyses.** Here, (1) a hypothetical genomic region is shown from the reference genotype genome and a sample representing another, different, genotype. The genomic region contains a dispersed duplication of a sequence (blue box) in the sample. The duplicated sequence contains a mutation, depicted with a red vertical line, relative to the Reference sequence. (2) Sequencing reads (depicted as coloured horizontal lines) are generated using short-read sequencing. When these sequencing reads are aligned to the reference genome (3), all reads from the duplicated sequence will align at a single location on the reference genome, with approximately half of the reads indicating a mutation, resulting in a heterozygous call. In case the dispersed duplication at its non-reference location affects the phenotype of interest, a genome-wide association study will instead indicate a significant association (depicted with a red dot above the red threshold line of significance) at the reference location, creating a characteristic single SNP peak, with no nearby associated SNPs in LD, which would be common for a regular association.

drawback of using T-DNA insertion lines is that they are typically in the Col-0 genetic background. Therefore, mutant analysis using T-DNA insertion lines relies on the assumption that Col-0 has a functional gene copy that is disrupted in the T-DNA insertion line, but this assumption does not always hold true. Additionally, T-DNA insertion lines may contain unknown additional T-DNA insertion copies and large-scale genomic rearrangements (Pucker *et al*., 2021), making it challenging to attribute the phenotypic effect solely to gene of interest without further confirmation. Nowadays, CRISPR Cas9-mediated knock-out mutants are the preferred alternative. Such mutants can be generated in any genetic background, while the effects of off-target mutations are likely reduced compared to T-DNA insertion lines. Off-target mutations induced by CRISPR-Cas9 have been reported at frequencies ranging from 10% up to 97% in *A. thaliana* but can be minimized by designing guide RNAs with high specificity (W. Xu *et al*., 2019). Especially by testing multiple independent mutants, the impact of potential off-target mutations is minimal.

Since mutant analysis alone is not sufficient to determine causality, it should be followed by genetic complementation of mutants using contrasting alleles of the candidate gene(s). Genetic complementation can be achieved through transgenic approaches, where contrasting natural alleles are cloned and transformed into a knock-out mutant background, or through natural crosses in a quantitative complementation, as demonstrated in **Chapter 4** for the validation of *FSD3*. In both methods, complementation of the mutant should be differentially affected by the contrasting alleles to establish causality. Transformation is the preferred method to achieve genetic complementation. However, cloning relatively large genomic fragments can be challenging and thus time-consuming. On the other hand, quantitative complementation is relatively straightforward, as it involves making crosses which is technically less demanding. Quantitative complementation relies on testing alleles in a hemizygous state, by crossing homozygous parents with contrasting alleles to a homozygous knock-out mutant (Mackay, 2001; Weigel, 2012). The resulting F_1 progeny, which differ at the hemizygous allele for the gene of interest, can then be compared. However, a limitation of this method is that the F_1 progeny are heterozygous for all loci that differ between the parents, which may influence the phenotypic expression. To account for heterosis effects in quantitative complementation, crosses between the same accessions with contrasting alleles and the genetic background of the knock-out mutant are performed. Multiple accessions for each allele should be used in these crosses to observe the allelic effects, which may thus necessitate many crosses.

6

Concluding remarks on GWAS

GWAS has both its strengths and limitations in its ability to detect genetic variants associated with a phenotype. The statistical power of GWAS depends on how much variance is generated by the causal variant, which depends on the phenotypic effect size contributed by an allele and the allele frequency (Visscher & Goddard, 2019). GWAS is particularly effective in detecting additive genetic effects, preferably with strong impact on the phenotype, especially when these variants are relatively common in the population. Genetic variants contributing to small effect-sizes, those that affect a phenotype in an epistatic manner, rare variants and variants under strong positive selection (especially those that are still largely in their original genetic background) are all examples of variants that are more challenging to detect through GWAS. Therefore, GWAS will be unlikely to fully dissolve the entire genetic basis of adaptation to the environment. Despite these limitations, the knowledge of genes discovered through GWAS, especially those associated with relatively simple additive genetic variation and that have been appropriately validated, can be highly valuable for agricultural applications. Allelic variants contributing to favourable traits have been tested in diverse genetic backgrounds and have typically demonstrated consistent phenotypic effects in natural environments. Incorporating these variants into elite cultivars holds promise for achieving desired phenotypic outcomes, provided they are successfully tested in the relevant specific environments.

Genome-environment association analysis to study signs of local adaptation

In **Chapter 4**, besides GWA analysis, I also conducted genome-environment association (GEA) analysis. Both association approaches relied on the same statistical methodology for the association analysis, making the statistical considerations described above equally applicable to GEA analysis. In GEA analysis, the goal is to assess whether genomic loci are associated with variation for environmental variables. Based on the assumption that local selection leads to an increases in local allele frequency, such an association indicates that the locus is involved in local adaptation (Rellstab *et al*., 2015). The majority of GEA studies in *A. thaliana* evaluated relatively large geographical regions (e.g. (Ferrero-Serrano & Assmann, 2019)), and typically indicate that climate adaptation is highly polygenic (Bay *et al*., 2017). Only very few studies investigated environmental adaptation in a relatively small region, most notably the study conducted in the south-west of France (Frachon *et al*., 2018), which already indicated climate adaptation as an important driver of local adaptation. Although the area in south-west France is comparable in size to the Netherlands, it does have a more diverse climate as it is under the influence of three contrasting climates. Nevertheless, our results also indicate that even relatively mild climatic clines can be important for shaping the genomic variation.

Although our results suggest signals of local adaptation, geographic and demographic processes can shape similar allelic distribution patterns and thereby mimic the patterns caused by adaptation (Rellstab *et al*., 2015). To distinguish between local adaptation and geographic and demographic processes, reciprocal transplant experiments can be very effective (de Villemereuil *et al*., 2016). In reciprocal transplant experiments, one tests whether the average fitness of locally adapted plants is higher than that of plants adapted in a different environment. However, for this to be effective, it relies on the assumption that the local selective forces that have shaped the allelic distributions are also predominantly at play while the reciprocal transplant experiment is conducted. When studying plants originating from highly contrasting climates, such an assumption will often be met. However, in the case of relatively mild and uniform climatic clines, such as in **Chapter 4**, allelic patterns may be mostly shaped by strong selection acting in years with relatively extreme conditions that can be relatively few and far apart. After all, allelic distributions that are detected in the present represent the evolutionary dynamics of the past.

Instead of using such reciprocal transplant experiments, in **Chapter 4** I tested whether allelic variation for the associated locus would differentially affect the phenotypic response to drought stress. Establishing that allelic variation actually has a phenotypic impact is important to establish whether it may contribute to local adaptation. The choice for drought was not straightforward, as environmental variables (precipitation and temperature variables) are highly correlated to another, and also likely to many other (a) biotic variables that were not considered. These correlations thus prevent from directly inferring which specific environmental variable(s) may have been drivers of (possible) local adaptation. If no differential phenotypic response would be observed for accessions with allelic differences for the associated locus, that would not disprove local adaptation, as one may not have tested the correct environmental variable(s). Similarly, in case a differential phenotypic response is observed, this does suggest that allelic variation affects the trait but does not demonstrate that this is relevant to local adaptation. However, the benefit of using a more controlled environment (and stress) is that it allows further testing of the effects of individual genes and thus to investigate how specific alleles affect the response to specific environmental variables. Therefore, the integration of this approach with reciprocal transplant experiments is most suitable in establishing whether a trait confers local adaptation and to investigate the genetic basis of it.

The high photosynthetic trait of *Hirschfeldia incana*

Studying the genomic basis of plant adaptation in wild species can provide highly relevant information that can be applied in agriculture, e.g. to increase crop yield. Yield increases can broadly be achieved either by increasing, and realizing, the yield potential and/or by decreasing the yield losses that may result from (a)biotic stresses. The very high levels of photosynthesis reported in **Chapter 5** for *H. incana* represent a possible case where both types of possible yield increases may meet. Firstly, it appears that *H. incana* is better adapted to high levels of irradiance. Although light is essential to drive photosynthesis, at high irradiance the absorbed light energy may exceed the energy that is used to drive photosynthesis or can be dissipated via photoprotective mechanisms (Demmig-Adams & Adams, 1992). Excessive absorbed light energy may then result in higher levels of free radicals that cause oxidative stress, mainly to photosystem II, and thereby can inhibit the plants photosynthetic capacity (Kato *et al*., 2003; Powles, 1984). *H. incana* especially outperforms the other tested species at higher irradiances (above 500 µmol · m^{-2} · s^{-1}) in terms of higher CO_2

assimilation levels (Figure 5.1). This suggests that the molecular and physiological photosynthetic machinery is better adapted to cope with high levels of irradiance and is thus less impaired by high light stress. Understanding the adaptations underlying this increased photosynthetic performance at high light also has the potential to transfer or optimise such traits in related crop species. This may thereby alleviate high light stress responses in such species, and, as will be discussed further below, may increase the yield potential.

In the recent years there has been a considerable interest in trying to improve photosynthesis as a means to improve crop yields (Evans, 2013; Theeuwen *et al*., 2022; van Bezouw *et al*., 2019; X.-G. Zhu *et al*., 2010). The rationale behind this approach seems straightforward, as biomass is to a large degree obtained from photosynthesis via the fixation of CO2 (Evans & von Caemmerer, 2011; Ort *et al*., 2015). Therefore, if photosynthesis can be made more efficient this should increase carbon resources, that may subsequently be redirected into the plant parts that are considered to contribute to crop yield. To a large degree, this rationale is supported by the yield potential model by Monteith, (1977) that suggests that photosynthesis is the key remaining factor that still has ample room for improvement in plants (Long *et al*., 2006). This model describes that yield potential (Y_p) is the product of the incoming photosynthetically active radiation and the efficiencies at which that radiation is intercepted ($ε_i$), is then converted to biomass ($ε_c$), and finally how biomass is partitioned into harvested product (n, also known as the harvest index) (Long *et al*., 2006). Both $ε_i$ and n have already been optimised close to their theoretical maximum, while there is ample room for improvement of $ε_c$, which is argued to be mostly dependent on the photosynthetic rates

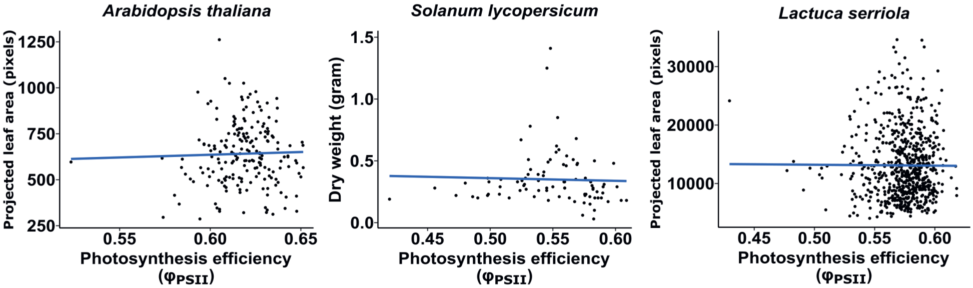


Figure 5: **Correlation analysis between photosynthesis efficiency and measures of growth (projected leaf area and dry weight).** All data shown here is gathered from plants that were grown under their respective control conditions, in a high-throughput phenotyping approach highly similar as is described for phenotyping the iron deficiency response of the DartMap in Chapter 4. The blue line represents a linear regression fitted through the data. For none of the three species tested a significant (positive) correlation is observed that were to be expected if improved photosynthesis would directly improve plant growth or yield. The data obtained from *Solanum lycopersicum* and *Lactuca serriola* were kindly provided by Laavanya Rayaprolu and Alan Pauls respectively.

(Long *et al*., 2006). ε_c is currently reported to be around 0.024 in C_3 crops but has a theoretical maximum of about 0.051 (Long *et al*., 2006). Taken together, there is broad consensus among photosynthesis scientists that enhanced photosynthesis is critical for enhancing crop yield (Garcia *et al*., 2023; Long *et al*., 2015; Simkin *et al*., 2019; Walter & Kromdijk, 2022; Wu *et al*., 2019).

Despite the seemingly overwhelming consensus among photosynthesis researchers on the importance of improving photosynthesis to increase crop yield, there are several observations that do not align with a simple relation between photosynthesis and yield (or biomass). For example, own observations during high-throughput phenotyping experiments of the DartMap panel and those of colleagues on natural variation panels in *A. thaliana* (e.g. HapMap and RegMap), lettuce (*Lactuca serriola* and *L. sativa*) and tomato, consistently reveal considerable natural variation for photosynthetic parameters, but also consistently demonstrate a lack of correlation between these photosynthetic parameters and plant growth (Figure 5).

Arguably, such a lack of correlation might be because these measurements are based on chlorophyl fluorescence (Maxwell & Johnson, 2000) and only represent components of the photosynthetic machinery that are used as a good proxy for photosynthesis (the light reactions) rather than the actual CO_2 assimilation rates (the dark reactions). Nevertheless, no convincing correlations are reported for CO_2 assimilation rates and yield within diversity panels in crops either (Driever *et al*., 2014; Koester *et al*., 2016; Sinclair *et al*., 2019). Moreover, as plant breeders have made tremendous yield increases in the past, a logical hypothesis would be that at least some of this has been achieved via (in)directly selecting for improved photosynthesis. That should thus reflect as a positive correlation between photosynthetic rates and yield when comparing elite cultivars to wild relatives. Nevertheless, again in practice there is little to no evidence for such correlations (Acevedo-Siaca *et al*., 2020; McAusland *et al*., 2020; Sinclair *et al*., 2019; Theeuwen *et al*., 2022). Such observations are especially odd in regards to the earlier mentioned claim by Long *et al*. (2006) that except for ε_c all other efficiencies have already been optimised to the near theoretical limit. Simple mathematics would then thus dictate that all further increases in yield potential obtained from that point should have predominantly been obtained via improving ε_c, which is mainly determined by photosynthesis. Overall, such consistent lack of correlations between rates of photosynthesis and yield across and within species at minimum shows that the relationship between these components is, at least in my opinion, far from as obvious and straightforward as is often stated.

What seems to complicate matters, is that it is not straightforward what improved photosynthesis really constitutes. For example, in the recent few years there has been especial attention to the photosynthetic response to fluctuating light conditions (Kaiser *et al*., 2018; Ruban, 2017; Slattery *et al*., 2018). As discussed before, plants have photoprotective mechanisms in place that can effectively deal with high light irradiance to prevent photoinhibition. However, when irradiance levels change from high to low, for instance because of a cloud moving in, such photoprotective mechanisms can be overprotective and restrict optimal use of the remaining light to drive photosynthesis (Ruban, 2017). A proposed solution to make the most out of the available light energy is to accelerate the recovery of photoprotective mechanisms. Kromdijk *et al*. (2016) achieved such faster light adaptation in tobacco by overexpressing components of the xanthophyll cycle and increasing the amounts of a PSII subunit. Their approach accelerated photoprotection, increased leaf CO_2 uptake, and more importantly increased dry matter production by about 15% in fluctuating light conditions during field trials. At first glance, this presents a clear-cut example of how improving a component of photosynthesis translates to higher yield. De Souza *et al*. (2022) applied the same strategy in soybean and reported that their method was also successful in improving photosynthesis and subsequently yield increased as well in some of their transgenic lines in one of their field trials, but not in their second season. In contrast, the same approach had the same impact on photosynthesis in *A. thaliana* but this impaired growth in fluctuating conditions (Garcia-Molina & Leister, 2020). These discrepancies in yield response exemplify that it is not obvious what constitutes improved photosynthesis, as the same physiological change can have completely opposing effects on growth and also appears to be dependent on the environment (i.e. the clear difference in two growing seasons observed by De Souza *et al*. (2022)). Regardless of the possible explanations on why these species may differ in their responses, much more important is the notion that a correlation between accelerated photoprotection and yield does at no instance prove causation and requires careful interpretation. As pointed out by Kaiser *et al*. (2019), the upregulated components of the xanthophyll cycle are also tightly integrated into the metabolic networks that contribute to the production of the hormone abscisic acid. So whether the observed changes in biomass are actually caused by accelerated photoprotection or result as a byproduct of other physiological processes that are affected as a consequence of the targeted xanthophyll cycle remains to be investigated.

Overall, my intention is not to claim that increasing photosynthesis cannot be a viable route towards improving crop yield potential. What the recent few decades have made clear is that photosynthesis is heavily integrated into the entirety of plant physiology and thus any single adjustment to photosynthetic components has in most (if not all) instances relationships to other plant components via feedback loops that make the relationship between measures of photosynthesis (capacity) unpredictable (Araus *et al*., 2021). Moreover, studying photosynthesis at the leaf level may not be the most effective as it does often not reflect well what happens at the canopy level (Araus *et al*., 2021). It will be important to better understand where the main rate limiting steps related to plant productivity are, as currently it is heavily debated whether plants are even carbon limited (Sinclair *et al*., 2019). Independent of the outcome of that discussion, it seems reasonable to assume that increasing carbon uptake is of little use when this is not coupled to increasing the uptake of other vital nutrients such as nitrogen and to increasing the phloem loading capacity of carbohydrates

(Araus *et al*., 2021; Sinclair *et al*., 2019). However, I would like to advocate for a more cautionary and critical view on claims such as '*Feeding the world: improving photosynthetic efficiency for sustainable crop production*' (Simkin *et al*., 2019) as long as there are no elite lines grown outside by farmers that actually live up to this promise despite the several decades of research in this direction. Moreover, I find it dubious and misleading how often studies such as the ones by Kromdijk *et al*. (2016) and De Souza *et al*. (2022) are cited as examples of how improved photosynthesis is causal to increased yields without any further reservations. To tackle the many challenges that future global food production faces, such as the need for higher productivity whilst dealing with the adverse effects of climate change, every promising avenue to achieve this should be explored. But as discussed in this thesis, applying a strong force of selection can most definitely speed up adaptation. There is no reason to doubt that this would not apply in shaping the ideas on what optimal photosynthesis constitutes and thus to address what the true potential of improved photosynthesis is to crop productivity.

Concluding remarks

The work presented in this thesis aimed to investigate the genetic basis of plant adaptation. Providing farmers and growers with crop varieties that are better adapted to their environment than current varieties, will continue to be a crucial challenge to meet agricultural demands. With rapid climate change, there will be increasingly more mismatches between current crop varieties and their desired performance in the field. Therefore, learning from the solutions that nature has evolved in response to a broad range of dynamic environmental factors will be essential in developing crops that are more tolerant to these conditions. Utilizing the existing natural variation found within individuals of species, as well as between closely related species (e.g. crops and their close wild relatives), will be vital for achieving this goal. The accessibility to, and quality of, whole genome DNA sequencing has significantly improved in the last decade, which rapidly reveals the large extent of genomic variability that exists among individuals of the same species. A growing number of high-quality long-read genome assemblies become available, while advances in phenomics are making it increasingly possible to associate these genetic variants with their phenotypic effects.

This work has also demonstrated that plant genetic adaptation to environment can occur rapidly, particularly in highly selective and stressful environments. The generation of genetic variation through mutation can play a significant role in rapidly contributing to producing better-adapted plants. A single large-effect SNP was sufficient to make a significant leap forward in zinc tolerance. However, not all traits may be equally amendable to this approach, as the genetic complexity underlying the trait will influence the likelihood of success. Traits for which mutations in single genes can directly and substantially impact the phenotype, such as disease resistance, are the most promising candidates for this approach. Harnessing the power of adaptive evolution to produce better-adapted plants is an interesting avenue to further explore in crops.

References

Acevedo-Siaca, L. G., Coe, R., Wang, Y., Kromdijk, J., Quick, W. P., & Long, S. P. (2020). Variation in photosynthetic induction between rice accessions and its potential for improving productivity. *New Phytologist*, *227*(4), 1097–1108.

Aguilera, A., & García-Muse, T. (2013). Causes of genome instability. *Annual Review of Genetics*, *47*, 1–32.

Airoldi, C. A., & Davies, B. (2012). Gene duplication and the evolution of plant MADS-box transcription factors. *Journal of Genetics and Genomics*, *39*(4), 157–165.

Akalin, A., Kormaksson, M., Li, S., Garrett-Bakelman, F. E., Figueroa, M. E., Melnick, A., & Mason, C. E. (2012). methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biology*, *13*(10), 1–9.

Alkan, C., Coe, B. P., & Eichler, E. E. (2011). Genome structural variation discovery and genotyping. *Nature Reviews Genetics*, *12*(5), 363–376.

Alonge, M., Wang, X., Benoit, M., Soyk, S., Pereira, L., Zhang, L., Suresh, H., Ramakrishnan, S., Maumus, F., & Ciren, D. (2020). Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell*, *182*(1), 145–161.

Alonso-Blanco, C., Andrade, J., Becker, C., Bemm, F., Bergelson, J., Borgwardt, K. M., Cao, J., Chae, E., Dezwaan, T. M., & Ding, W. (2016). 1,135 genomes reveal the global pattern of polymorphism in Arabidopsis thaliana. *Cell*, *166*(2), 481–491.

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*, 5743–5754.

Alvarez Prado, S., Sanchez, I., Cabrera-Bosquet, L., Grau, A., Welcker, C., Tardieu, F., & Hilgert, N. (2019). To clean or not to clean phenotypic datasets for outlier plants in genetic analyses? *Journal of Experimental Botany*, *70*(15), 3693–3698.

Anderson, J. K., & Warwick, S. I. (1999). Chromosome number evolution in the tribe Brassiceae (Brassicaceae): evidence from isozyme number. *Plant Systematics and Evolution*, *215*, 255–285.

Andrés, F., & Coupland, G. (2012). The genetic basis of flowering responses to seasonal cues. *Nature Reviews Genetics*, *13*(9), 627–639.

Araus, J. L., Sanchez-Bragado, R., & Vicente, R. (2021). Improving crop yield and resilience through optimization of photosynthesis: panacea or pipe dream? *Journal of Experimental Botany*, *72*(11), 3936–3955.

Ariani, A., Berny Mier y Teran, J. C., & Gepts, P. (2016). Genome-wide identification of SNPs and copy number variation in common bean (Phaseolus vulgaris L.) using genotyping-by-sequencing (GBS). *Molecular Breeding*, *36*, 1–11.

Arias, T., & Pires, J. C. (2012). A fully resolved chloroplast phylogeny of the brassica crops and wild relatives (Brassicaceae: Brassiceae): Novel clades and potential taxonomic implications. *Taxon*, *61*(5), 980–988.

Arora, R., Agarwal, P., Ray, S., Singh, A. K., Singh, V. P., Tyagi, A. K., & Kapoor, S. (2007). MADS-box gene family in rice: genome-wide identification, organization and expression profiling during reproductive development and stress. *BMC Genomics*, *8*(1), 1–21.

Arumuganathan, K., & Earle, E. D. (1991). Estimation of nuclear DNA content of plants by flow cytometry. *Plant Molecular Biology Reporter*, *9*, 229–241.

Atwell, S., Huang, Y. S., Vilhjálmsson, B. J., Willems, G., Horton, M., Li, Y., Meng, D., Platt, A., Tarone, A. M., & Hu, T. T. (2010). Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature*, *465*(7298), 627–631.

Auguy, F., Fahr, M., Moulin, P., Brugel, A., Laplaze, L., Mzibri, M. El, Filali-Maltouf, A., Doumas, P., & Smouni, A. (2013). Lead tolerance and accumulation in Hirschfeldia incana, a Mediterranean Brassicaceae from metalliferous mine spoils. *PLoS One*, *8*(5), e61932.

Auguy, F., Fahr, M., Moulin, P., El Mzibri, M., Smouni, A., Filali-Maltouf, A., Béna, G., & Doumas, P. (2016). Transcriptome changes in Hirschfeldia incana in response to lead exposure. *Frontiers in Plant Science*, *6*, 1231.

Aybar, C., Wu, Q., Bautista, L., Yali, R., & Barja, A. (2020). rgee: An R package for interacting with Google Earth Engine. *Journal of Open Source Software*, *5*(51), 2272.

Baduel, P., Leduque, B., Ignace, A., Gy, I., Gil Jr, J., Loudet, O., Colot, V., & Quadrana, L. (2021). Genetic and environmental modulation of transposition shapes the evolutionary potential of Arabidopsis thaliana. *Genome Biology*, *22*(1), 138.

Bagheri, H., El-Soda, M., van Oorschot, I., Hanhart, C., Bonnema, G., Jansen-van den Bosch, T., Mank, R., Keurentjes, J. J. B., Meng, L., & Wu, J. (2012). Genetic analysis of morphological traits in a new, versatile, rapid-cycling Brassica rapa recombinant inbred line population. *Frontiers in Plant Science*, *3*, 183.

Bai, Z., Chen, J., Liao, Y., Wang, M., Liu, R., Ge, S., Wing, R. A., & Chen, M. (2016). The impact and origin of copy number variations in the Oryza species. *BMC Genomics*, *17*, 1–12.

Baker, N. R., & Oxborough, K. (2004). Chlorophyll fluorescence as a probe of photosynthetic productivity. In *Chlorophyll a fluorescence: a signature of photosynthesis* (pp. 65–82). Springer.

Bancroft, I., Morgan, C., Fraser, F., Higgins, J., Wells, R., Clissold, L., Baker, D., Long, Y., Meng, J., & Wang, X. (2011). Dissecting the genome of the polyploid crop oilseed rape by transcriptome sequencing. *Nature Biotechnology*, *29*(8), 762–766.

Bank, C., Ewing, G. B., Ferrer-Admettla, A., Foll, M., & Jensen, J. D. (2014). Thinking too positive? Revisiting current methods of population genetic selection inference. *Trends in Genetics*, *30*(12), 540–546.

Barboza, L., Effgen, S., Alonso-Blanco, C., Kooke, R., Keurentjes, J. J. B., Koornneef, M., & Alcázar, R. (2013). Arabidopsis semidwarfs evolved from independent mutations in GA20ox1, ortholog to green revolution dwarf alleles in rice and barley. *Proceedings of the National Academy of Sciences*, *110*(39), 15818–15823.

Barrick, J. E., & Lenski, R. E. (2013). Genome dynamics during experimental evolution. *Nature Reviews Genetics*, *14*(12), 827–839.

Bartoli, C., Frachon, L., Barret, M., Rigal, M., Huard-Chauveau, C., Mayjonade, B., Zanchetta, C., Bouchez, O., Roby, D., & Carrère, S. (2018). In situ relationships between microbiota and potential pathobiota in Arabidopsis thaliana. *The ISME Journal*, *12*(8), 2024–2038.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *ArXiv Preprint ArXiv:1406.5823*.

Baxter, A., Mittler, R., & Suzuki, N. (2014). ROS as key players in plant stress signalling. *Journal of Experimental Botany*, *65*(5), 1229–1240.

Baxter, I., Brazelton, J. N., Yu, D., Huang, Y. S., Lahner, B., Yakubova, E., Li, Y., Bergelson, J., Borevitz, J. O., & Nordborg, M. (2010). A coastal cline in sodium accumulation in Arabidopsis thaliana is driven by natural variation of the sodium transporter AtHKT1; 1. *PLoS Genetics*, *6*(11), e1001193.

Bay, R. A., Rose, N., Barrett, R., Bernatchez, L., Ghalambor, C. K., Lasky, J. R., Brem, R. B., Palumbi, S. R., & Ralph, P. (2017). Predicting responses to contemporary environmental change using evolutionary response architectures. *The American Naturalist*, *189*(5), 463–473.

Bayfield, M. A., Yang, R., & Maraia, R. J. (2010). Conserved and divergent features of the structure and function of La and La-related proteins (LARPs). *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, *1799*(5–6), 365–378.

Beilstein, M. A., Nagalingum, N. S., Clements, M. D., Manchester, S. R., & Mathews, S. (2010). Dated molecular phylogenies indicate a Miocene origin for Arabidopsis thaliana. *Proceedings of the National Academy of Sciences*, *107*(43), 18724–18728.

Belfield, E. J., Brown, C., Ding, Z. J., Chapman, L., Luo, M., Hinde, E., Van Es, S. W., Johnson, S., Ning, Y., & Zheng, S. J. (2021). Thermal stress accelerates Arabidopsis thaliana mutation rate. *Genome Research*, *31*(1), 40–50.

Beló, A., Beatty, M. K., Hondred, D., Fengler, K. A., Li, B., & Rafalski, A. (2010). Allelic genome structural variations in maize detected by array comparative genome hybridization. *Theoretical and Applied Genetics*, *120*(2), 355–367.

Belser, C., Istace, B., Denis, E., Dubarry, M., Baurens, F.-C., Falentin, C., Genete, M., Berrabah, W., Chèvre, A.-M., & Delourme, R. (2018). Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. *Nature Plants*, *4*(11), 879–887.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300.

Bennett, M. D., Leitch, I. J., Price, H. J., & Johnston, J. S. (2003). Comparisons with Caenorhabditis (~ 100 Mb) and Drosophila (~ 175 Mb) using flow cytometry show genome size in Arabidopsis to be ~157 Mb and thus ~25% larger than the Arabidopsis genome initiative estimate of ~125 Mb. *Annals of Botany*, *91*(5), 547–557.

Beric, A., Mabry, M. E., Harkess, A. E., Brose, J., Schranz, M. E., Conant, G. C., Edger, P. P., Meyers, B. C., & Pires, J. C. (2021). Comparative phylogenetics of repetitive elements in a diverse order of flowering plants (Brassicales). *G3*, *11*(7), jkab140.

Bewick, A. J., & Schmitz, R. J. (2017). Gene body DNA methylation in plants. *Current Opinion in Plant Biology*, *36*, 103–110.

Bienert, G. P., Møller, A. L. B., Kristiansen, K. A., Schulz, A., Møller, I. M., Schjoerring, J. K., & Jahn, T. P. (2007). Specific aquaporins facilitate the diffusion of hydrogen peroxide across membranes. *Journal of Biological Chemistry*, *282*(2), 1183–1192.

Billings, W. D., & Mooney, H. A. (1968). The ecology of arctic and alpine plants. *Biological Reviews*, *43*(4), 481–529.

Boocock, J., Chagné, D., Merriman, T. R., & Black, M. A. (2015). The distribution and impact of common copy-number variation in the genome of the domesticated apple, Malus x domestica Borkh. *BMC Genomics*, *16*(1), 1–15.

Boyko, A., Hudson, D., Bhomkar, P., Kathiria, P., & Kovalchuk, I. (2006). Increase of homologous recombination frequency in vascular tissue of Arabidopsis plants exposed to salt stress. *Plant and Cell Physiology*, *47*(6), 736–742.

Brachi, B., Villoutreix, R., Faure, N., Hautekèete, N., Piquot, Y., Pauwels, M., Roby, D., Cuguen, J., Bergelson, J., & Roux, F. (2013). Investigation of the geographical scale of adaptive phenological variation and its underlying genetics in A rabidopsis thaliana. *Molecular Ecology*, *22*(16), 4222–4240.

Broadley, M. R., White, P. J., Hammond, J. P., Zelko, I., & Lux, A. (2007). Zinc in plants. *New Phytologist*, *173*(4), 677–702.

Bruce, T. J. A., Matthes, M. C., Napier, J. A., & Pickett, J. A. (2007). Stressful "memories" of plants: evidence and possible mechanisms. *Plant Science*, *173*(6), 603–608.

Brunner, S., Fengler, K., Morgante, M., Tingey, S., & Rafalski, A. (2005). Evolution of DNA sequence nonhomologies among maize inbreds. *The Plant Cell*, *17*(2), 343–360.

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, *12*(1), 59–60.

Cai, X., Wu, J., Liang, J., Lin, R., Zhang, K., Cheng, F., & Wang, X. (2020). Improved Brassica oleracea JZS assembly reveals significant changing of LTR-RT dynamics in different morphotypes. *Theoretical and Applied Genetics*, *133*, 3187–3199.

Campbell, C. D., & Eichler, E. E. (2013). Properties and rates of germline mutations in humans. *Trends in Genetics*, *29*(10), 575–584.

Campos, A. C. A. L., van Dijk, W. F. A., Ramakrishna, P., Giles, T., Korte, P., Douglas, A., Smith, P., & Salt, D. E. (2021). 1,135 ionomes reveal the global pattern of leaf and seed mineral nutrient and trace element diversity in Arabidopsis thaliana. *The Plant Journal*, *106*(2), 536–554.

Canvin, D. T., Berry, J. A., Badger, M. R., Fock, H., & Osmond, C. B. (1980). Oxygen exchange in leaves in the light. *Plant Physiology*, *66*(2), 302–307

Cao, J., Schneeberger, K., Ossowski, S., Günther, T., Bender, S., Fitz, J., Koenig, D., Lanz, C., Stegle, O., & Lippert, C. (2011). Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nature Genetics*, *43*(10), 956–963.

Cappa, J. J., & Pilon-Smits, E. A. H. (2014). Evolutionary aspects of elemental hyperaccumulation. *Planta*, *239*, 267–275.

Cardone, M. F., D'Addabbo, P., Alkan, C., Bergamini, C., Catacchio, C. R., Anaclerio, F., Chiatante, G., Marra, A., Giannuzzi, G., & Perniola, R. (2016). Inter-varietal structural variation in grapevine genomes. *The Plant Journal*, *88*(4), 648–661.

Carretero-Paulet, L., & Fares, M. A. (2012). Evolutionary dynamics and functional specialization of plant paralogs formed by whole and small-scale genome duplications. *Molecular Biology and Evolution*, *29*(11), 3541–3551.

Carvalho, C. M. B., & Lupski, J. R. (2016). Mechanisms underlying structural variant formation in genomic disorders. *Nature Reviews Genetics*, *17*(4), 224–238.

Castro, B., Citterico, M., Kimura, S., Stevens, D. M., Wrzaczek, M., & Coaker, G. (2021). Stress-induced reactive oxygen species compartmentalization, perception and signalling. *Nature Plants*, *7*(4), 403–412.

Chain, F. J. J., Flynn, J. M., Bull, J. K., & Cristescu, M. E. (2019). Accelerated rates of large-scale mutations in the presence of copper and nickel. *Genome Research*, *29*(1), 64–73.

Chalhoub, B., Denoeud, F., Liu, S., Parkin, I. A. P., Tang, H., Wang, X., Chiquet, J., Belcram, H., Tong, C., & Samans, B. (2014). Early allopolyploid evolution in the post-Neolithic Brassica napus oilseed genome. *Science*, *345*(6199), 950–953.

Chang, C., Lu, J., Zhang, H.-P., Ma, C.-X., & Sun, G. (2015). Copy number variation of cytokinin oxidase gene Tackx4 associated with grain weight and chlorophyll content of flag leaf in common wheat. *PLoS One*, *10*(12), e0145970.

Chang, S., Puryear, J., & Cairney, J. (1993). A simple and efficient method for isolating RNA from pine trees. *Plant Molecular Biology Reporter*, *11*, 113–116.

Chang, S.-L., Lai, H.-Y., Tung, S.-Y., & Leu, J.-Y. (2013). Dynamic large-scale chromosomal rearrangements fuel rapid adaptation in yeast populations. *PLoS Genetics*, *9*(1), e1003232.

Chia, J.-M., Song, C., Bradbury, P. J., Costich, D., De Leon, N., Doebley, J., Elshire, R. J., Gaut, B., Geller, L., & Glaubitz, J. C. (2012). Maize HapMap2 identifies extant variation from a genome in flux. *Nature Genetics*, *44*(7), 803–807.

Choi, S. R., Teakle, G. R., Plaha, P., Kim, J. H., Allender, C. J., Beynon, E., Piao, Z. Y., Soengas, P., Han, T. H., & King, G. J. (2007). The reference genetic linkage map for the multinational Brassica rapa genome sequencing project. *Theoretical and Applied Genetics*, *115*, 777–792.

Choudhury, F. K., Rivero, R. M., Blumwald, E., & Mittler, R. (2017). Reactive oxygen species, abiotic stress and stress combination. *The Plant Journal*, *90*(5), 856–867.

Collins, S., & De Meaux, J. (2009). Adaptation to different rates of environmental change in Chlamydomonas. *Evolution*, *63*(11), 2952–2965.

Conrad, T. M., Lewis, N. E., & Palsson, B. Ø. (2011). Microbial laboratory evolution in the era of genome-scale science. *Molecular Systems Biology*, *7*(1), 509.

Consortium, B. rapa G. S. P., Wang, X., Wang, H., Wang, J., Sun, R., Wu, J., Liu, S., Bai, Y., Mun, J.-H., & Bancroft, I. (2011). The genome of the mesopolyploid crop species Brassica rapa. *Nature Genetics*, *43*(10), 1035–1039.

Cook, D. E., Bayless, A. M., Wang, K., Guo, X., Song, Q., Jiang, J., & Bent, A. F. (2014). Distinct copy number, coding sequence, and locus methylation patterns underlie Rhg1-mediated soybean resistance to soybean cyst nematode. *Plant Physiology*, *165*(2), 630–647.

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. (2012). Copy number variation of multiple genes at Rhg1 mediates nematode resistance in soybean. *Science*, *338*(6111), 1206–1209.

Coombe, L., Zhang, J., Vandervalk, B. P., Chu, J., Jackman, S. D., Birol, I., & Warren, R. L. (2018). ARKS: chromosome-scale scaffolding of human genome drafts with linked read kmers. *BMC Bioinformatics*, *19*(1), 1–10.

Craciun, A. R., Meyer, C.-L., Chen, J., Roosens, N., De Groodt, R., Hilson, P., & Verbruggen, N. (2012). Variation in HMA4 gene copy number and expression among Noccaea caerulescens populations presenting different levels of Cd tolerance and accumulation. *Journal of Experimental Botany*, *63*(11), 4179–4189.

Crafts-Brandner, S. J., & Salvucci, M. E. (2002). Sensitivity of photosynthesis in a C4 plant, maize, to heat stress. *Plant Physiology*, *129*(4), 1773–1780.

Crocco, C. D., Ocampo, G. G., Ploschuk, E. L., Mantese, A., & Botto, J. F. (2018). Heterologous expression of AtBBX21 enhances the rate of photosynthesis and alleviates photoinhibition in Solanum tuberosum. *Plant Physiology*, *177*(1), 369–380.

Darmency, & Fleury. (2000). Mating system in Hirschfeldia incana and hybridization to oilseed rape. *Weed Research*, *40*(2), 231–238.

Dassanayake, M., Oh, D.-H., Haas, J. S., Hernandez, A., Hong, H., Ali, S., Yun, D.-J., Bressan, R. A., Zhu, J.-K., & Bohnert, H. J. (2011). The genome of the extremophile crucifer Thellungiella parvula. *Nature Genetics*, *43*(9), 913–918.

De Souza, A. P., Burgess, S. J., Doran, L., Hansen, J., Manukyan, L., Maryn, N., Gotarkar, D., Leonelli, L., Niyogi, K. K., & Long, S. P. (2022). Soybean photosynthesis and crop yield are improved by accelerating recovery from photoprotection. *Science*, *377*(6608), 851–854.

de Villemereuil, P., Gaggiotti, O. E., Mouterde, M., & Till-Bottraud, I. (2016). Common garden experiments in the genomic era: new perspectives and opportunities. *Heredity*, *116*(3), 249–254.

De Vries, F., De Groot, W. J. M., Hoogland, T., & Denneboom, J. (2003). *De Bodemkaart van Nederland digitaal; toelichting bij inhoud, actualiteit en methodiek en korte beschrijving van additionele informatie*. Alterra.

DeBolt, S. (2010). Copy number variation shapes genome diversity in Arabidopsis over immediate family generational scales. *Genome Biology and Evolution*, *2*, 441–453.

Defoort, J., Van de Peer, Y., & Carretero-Paulet, L. (2019). The evolution of gene duplicates in angiosperms and the impact of protein–protein interactions and the mechanism of duplication. *Genome Biology and Evolution*, *11*(8), 2292–2305.

Demmig-Adams, B., & Adams Iii, W. W. (1992). Photoprotection and other responses of plants to high light stress. *Annual Review of Plant Biology*, *43*(1), 599–626.

Díaz, A., Zikhali, M., Turner, A. S., Isaac, P., & Laurie, D. A. (2012). Copy number variation affecting the Photoperiod-B1 and Vernalization-A1 genes is associated with altered flowering time in wheat (Triticum aestivum). *PloS One*, *7*(3), e33234.

Dillon, A., Varanasi, V. K., Danilova, T. V, Koo, D.-H., Nakka, S., Peterson, D. E., Tranel, P. J., Friebe, B., Gill, B. S., & Jugulam, M. (2017). Physical mapping of amplified copies of the 5-enolpyruvylshikimate-3-phosphate synthase gene in glyphosate-resistant Amaranthus tuberculatus. *Plant Physiology*, *173*(2), 1226–1234.

Dolatabadian, A., Patel, D. A., Edwards, D., & Batley, J. (2017). Copy number variation and disease resistance in plants. *Theoretical and Applied Genetics*, *130*, 2479–2490.

Dräger, D. B., Desbrosses-Fonrouge, A., Krach, C., Chardonnens, A. N., Meyer, R. C., Saumitou-Laprade, P., & Krämer, U. (2004). Two genes encoding Arabidopsis halleri MTP1 metal transport proteins co-segregate with zinc tolerance and account for high MTP1 transcript levels. *The Plant Journal*, *39*(3), 425–439.

Driever, S. M., Lawson, T., Andralojc, P. J., Raines, C. A., & Parry, M. A. J. (2014). Natural variation in photosynthetic capacity, growth, and yield in 64 field-grown wheat genotypes. *Journal of Experimental Botany*, *65*(17), 4959–4973.

Du, J., Tian, Z., Sui, Y., Zhao, M., Song, Q., Cannon, S. B., Cregan, P., & Ma, J. (2012). Pericentromeric effects shape the patterns of divergence, retention, and expression of duplicated genes in the paleopolyploid soybean. *The Plant Cell*, *24*(1), 21–32.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797.

Edger, P. P., & Pires, J. C. (2009). Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes. *Chromosome Research*, *17*, 699–717.

Edwards, E. J., Osborne, C. P., Strömberg, C. A. E., Smith, S. A., Consortium, C. G., Bond, W. J., Christin, P.-A., Cousins, A. B., Duvall, M. R., & Fox, D. L. (2010). The origins of C4 grasslands: integrating evolutionary and ecosystem science. *Science*, *328*(5978), 587–591.

Ehleringer, J. (1985). Annuals and perennials of warm deserts. In *Physiological ecology of North American plant communities* (pp. 162–180). Springer.

Ellinghaus, D., Kurtz, S., & Willhoeft, U. (2008). LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics*, *9*, 1–14.

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloS One*, *6*(5), e19379.

Emms, D. M., & Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology*, *20*, 1–14.

Evans, J. R. (2013). Improving photosynthesis. *Plant Physiology*, *162*(4), 1780–1793.

Evans, J. R., & von Caemmerer, S. (2011). Enhancing photosynthesis. In *Plant physiology* (Vol. 155, Issue 1, p. 19). American Society of Plant Biologists.

Exposito-Alonso, M., Burbano, H. A., Bossdorf, O., Nielsen, R., & Weigel, D. (2019). Natural selection on the Arabidopsis thaliana genome in present and future climates. *Nature*, *573*(7772), 126–129.

Fahr, M., Laplaze, L., El Mzibri, M., Doumas, P., Bendaou, N., Hocher, V., Bogusz, D., & Smouni, A. (2015). Assessment of lead tolerance and accumulation in metallicolous and non-metallicolous populations of Hirschfeldia incana. *Environmental and Experimental Botany*, *109*, 186–192.

Fan, X., Chaisson, M., Nakhleh, L., & Chen, K. (2017). HySA: a Hybrid Structural variant Assembly approach using next-generation and single-molecule sequencing technologies. *Genome Research*, *27*(5), 793–800.

Favory, J.-J., Kobayshi, M., Tanaka, K., Peltier, G., Kreis, M., Valay, J.-G., & Lerbs-Mache, S. (2005). Specific function of a plastid sigma factor for ndh F gene transcription. *Nucleic Acids Research*, *33*(18), 5991–5999.

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.

Fichman, Y., & Mittler, R. (2020). Rapid systemic signaling during abiotic and biotic stresses: is the ROS wave master of all trades? *The Plant Journal*, *102*(5), 887–896.

Flood, P. J., Harbinson, J., & Aarts, M. G. M. (2011). Natural genetic variation in plant photosynthesis. *Trends in Plant Science*, *16*(6), 327–335.

Flood, P. J., Kruijer, W., Schnabel, S. K., van der Schoor, R., Jalink, H., Snel, J. F. H., Harbinson, J., & Aarts, M. G. M. (2016). Phenomics for photosynthesis, growth and reflectance in Arabidopsis thaliana reveals circadian and long-term fluctuations in heritability. *Plant Methods*, *12*(1), 1–14.

Flood, P. J., Van Heerwaarden, J., Becker, F., De Snoo, C. B., Harbinson, J., & Aarts, M. G. M. (2016). Whole-genome hitchhiking on an organelle mutation. *Current Biology*, *26*(10), 1306–1311.

Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., & Smit, A. F. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences*, *117*(17), 9451–9457.

Fonseca, S., Chico, J. M., & Solano, R. (2009). The jasmonate pathway: the ligand, the receptor and the core signalling module. *Current Opinion in Plant Biology*, *12*(5), 539–547.

Fournier-Level, A., Wilczek, A. M., Cooper, M. D., Roe, J. L., Anderson, J., Eaton, D., Moyers, B. T., Petipas, R. H., Schaeffer, R. N., & Pieper, B. (2013). Paths to selection on life history loci in different natural environments across the native range of A rabidopsis thaliana. *Molecular Ecology*, *22*(13), 3552–3566.

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.

Francia, E., Morcia, C., Pasquariello, M., Mazzamurro, V., Milc, J. A., Rizza, F., Terzi, V., & Pecchioni, N. (2016). Copy number variation at the HvCBF4–HvCBF2 genomic segment is a major component of frost resistance in barley. *Plant Molecular Biology*, *92*, 161–175.

Franzke, A., German, D., Al-Shehbaz, I. A., & Mummenhoff, K. (2009). Arabidopsis family ties: molecular phylogeny and age estimates in Brassicaceae. *Taxon*, *58*(2), 425–437.

Freeling, M., Scanlon, M. J., & Fowler, J. E. (2015). Fractionation and subfunctionalization following genome duplications: mechanisms that drive gene content and their consequences. *Current Opinion in Genetics & Development*, *35*, 110–118.

Friedberg, E. C., Walker, G. C., Siede, W., & Wood, R. D. (2005). *DNA repair and mutagenesis*. American Society for Microbiology Press.

Furbank, R. T., Jimenez-Berni, J. A., George-Jaeggli, B., Potgieter, A. B., & Deery, D. M. (2019). Field crop phenomics: enabling breeding for radiation use efficiency and biomass in cereal crops. *New Phytologist*, *223*(4), 1714–1727.

Furumizu, C., Alvarez, J. P., Sakakibara, K., & Bowman, J. L. (2015). Antagonistic roles for KNOX1 and KNOX2 genes in patterning the land plant body plan following an ancient gene duplication. *PLoS Genetics*, *11*(2), e1004980.

Gaines, T. A., Barker, A. L., Patterson, E. L., Westra, P., Westra, E. P., Wilson, R. G., Jha, P., Kumar, V., & Kniss, A. R. (2016). EPSPS gene copy number and whole-plant glyphosate resistance level in Kochia scoparia. *PLoS One*, *11*(12), e0168295.

Gaines, T. A., Shaner, D. L., Ward, S. M., Leach, J. E., Preston, C., & Westra, P. (2011). Mechanism of resistance of evolved glyphosate-resistant Palmer amaranth (Amaranthus palmeri). *Journal of Agricultural and Food Chemistry*, *59*(11), 5886–5889.

Gaines, T. A., Wright, A. A., Molin, W. T., Lorentz, L., Riggins, C. W., Tranel, P. J., Beffa, R., Westra, P., & Powles, S. B. (2013). Identification of genetic elements associated with EPSPS gene amplification. *PloS One*, *8*(6), e65819.

Gaines, T. A., Zhang, W., Wang, D., Bukun, B., Chisholm, S. T., Shaner, D. L., Nissen, S. J., Patzoldt, W. L., Tranel, P. J., & Culpepper, A. S. (2010). Gene amplification confers glyphosate resistance in Amaranthus palmeri. *Proceedings of the National Academy of Sciences*, *107*(3), 1029–1034.

Gaio, D., Anantanawat, K., To, J., Liu, M., Monahan, L., & Darling, A. E. (2022). Hackflex: low-cost, high-throughput, Illumina Nextera Flex library construction. *Microbial Genomics*, *8*(1).

Gallie, D. R., & Chen, Z. (2019). Chloroplast-localized iron superoxide dismutases FSD2 and FSD3 are functionally distinct in Arabidopsis. *PloS One*, *14*(7), e0220078.

Garcia, A., Gaju, O., Bowerman, A. F., Buck, S. A., Evans, J. R., Furbank, R. T., Gilliham, M., Millar, A. H., Pogson, B. J., & Reynolds, M. P. (2023). Enhancing crop yields through improvements in the efficiency of photosynthesis and respiration. *New Phytologist*, *237*(1), 60–77.

Garcia-Molina, A., & Leister, D. (2020). Accelerated relaxation of photoprotection impairs biomass accumulation in Arabidopsis. *Nature Plants*, *6*(1), 9–12.

Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *ArXiv Preprint ArXiv:1207.3907*.

Genty, B., & Harbinson, J. (1996). Regulation of light utilization for photosynthetic electron transport. *Photosynthesis and the Environment*, 67–99.

Gibson, A. C. (1998). Photosynthetic organs of desert plants. *Bioscience*, *48*(11), 911–920.

Gilbert, M. E., Zwieniecki, M. A., & Holbrook, N. M. (2011). Independent variation in photosynthetic capacity and stomatal conductance leads to differences in intrinsic water use efficiency in 11 soybean genotypes before and during mild drought. *Journal of Experimental Botany*, *62*(8), 2875–2887.

Gitelson, A. A., Peng, Y., Arkebauer, T. J., & Suyker, A. E. (2015). Productivity, absorbed photosynthetically active radiation, and light use efficiency in crops: Implications for remote sensing of crop primary production. *Journal of Plant Physiology*, *177*, 100–109.

Göktay, M., Fulgione, A., & Hancock, A. M. (2021). A new catalog of structural variants in 1,301 A. thaliana lines from Africa, Eurasia, and North America reveals a signature of balancing selection at defense response genes. *Molecular Biology and Evolution*, *38*(4), 1498–1511.

Golicz, A. A., Bayer, P. E., Barker, G. C., Edger, P. P., Kim, H., Martinez, P. A., Chan, C. K. K., Severn-Ellis, A., McCombie, W. R., & Parkin, I. A. P. (2016). The pangenome of an agronomically important crop plant Brassica oleracea. *Nature Communications*, *7*(1), 13390.

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, *17*(6), 333–351. https://doi.org/10.1038/nrg.2016.49

Gorter, F. A., Derks, M. F. L., van den Heuvel, J., Aarts, M. G. M., Zwaan, B. J., de Ridder, D., & de Visser, J. A. G. M. (2017). Genomics of adaptation depends on the rate of environmental change in experimental yeast populations. *Molecular Biology and Evolution*, *34*(10), 2613–2626.

Grace, J. (1988). 3. Plant response to wind. *Agriculture, Ecosystems & Environment*, *22*, 71–88.
Grandbastien, M.-A. (2015). LTR retrotransposons, handy hitchhikers of plant regulation and stress response. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, *1849*(4), 403–416.

Gu, C., Wang, L., Wang, W., Zhou, H., Ma, B., Zheng, H., Fang, T., Ogutu, C., Vimolmangkang, S., & Han, Y. (2016). Copy number variation of a gene cluster encoding endopolygalacturonase mediates flesh texture and stone adhesion in peach. *Journal of Experimental Botany*, *67*(6), 1993–2005.

Gu, J., Yin, X., Stomph, T., & Struik, P. C. (2014). Can exploiting natural genetic variation in leaf photosynthesis contribute to increasing rice productivity? A simulation analysis. *Plant, Cell & Environment*, *37*(1), 22–34.

Gu, J., Yin, X., Struik, P. C., Stomph, T. J., & Wang, H. (2012). Using chromosome introgression lines to map quantitative trait loci for photosynthesis parameters in rice (Oryza sativa L.) leaves under drought and well-watered field conditions. *Journal of Experimental Botany*, *63*(1), 455–469.

Gu, J., Zhou, Z., Li, Z., Chen, Y., Wang, Z., Zhang, H., & Yang, J. (2017). Photosynthetic properties and potentials for improvement of photosynthesis in pale green leaf rice under high light conditions. *Frontiers in Plant Science*, *8*, 1082.

Gu, W., Zhang, F., & Lupski, J. R. (2008). Mechanisms for human genomic rearrangements. *Pathogenetics*, *1*, 1–17.

Guan, D., McCarthy, S. A., Wood, J., Howe, K., Wang, Y., & Durbin, R. (2020). Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics*, *36*(9), 2896–2898.

Gupta, P. K., Kulwal, P. L., & Jaiswal, V. (2019). Association mapping in plants in the post-GWAS genomics era. *Advances in Genetics*, *104*, 75–154.

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, *29*(8), 1072–1075.

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., White, O., Buell, C. R., & Wortman, J. R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biology*, *9*, 1–22.

Halligan, D. L., & Keightley, P. D. (2009). Spontaneous mutation accumulation studies in evolutionary genetics. *Annual Review of Ecology, Evolution, and Systematics*, *40*, 151–172.

Han, B., Yang, Z., Samma, M. K., Wang, R., & Shen, W. (2013). Systematic validation of candidate reference genes for qRT-PCR normalization under iron deficiency in Arabidopsis. *Biometals*, *26*, 403–413.

Han, J., Köster, P., Drerup, M. M., Scholz, M., Li, S., Edel, K. H., Hashimoto, K., Kuchitsu, K., Hippler, M., & Kudla, J. (2019). Fine-tuning of RBOHF activity is achieved by differential phosphorylation and Ca2+ binding. *New Phytologist*, *221*(4), 1935–1949.

Hanada, K., Vallejo, V., Nobuta, K., Slotkin, R. K., Lisch, D., Meyers, B. C., Shiu, S.-H., & Jiang, N. (2009). The functional role of pack-MULEs in rice inferred from purifying selection and expression profile. *The Plant Cell*, *21*(1), 25–38.

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.

Hanikenne, M., Talke, I. N., Haydon, M. J., Lanz, C., Nolte, A., Motte, P., Kroymann, J., Weigel, D., & Krämer, U. (2008). Evolution of metal hyperaccumulation required cis-regulatory changes and triplication of HMA4. *Nature*, *453*(7193), 391–395.

Hänsch, R., & Mendel, R. R. (2009). Physiological functions of mineral micronutrients (cu, Zn, Mn, Fe, Ni, Mo, B, cl). *Current Opinion in Plant Biology*, *12*(3), 259–266.

Harbinson, J., Kaiser, E., & Morales, A. S. (2022). Integrating the stages of photosynthesis. In *Photosynthesis in action* (pp. 195–242). Elsevier.

Hardigan, M. A., Crisovan, E., Hamilton, J. P., Kim, J., Laimbeer, P., Leisner, C. P., Manrique-Carpintero, N. C., Newton, L., Pham, G. M., & Vaillancourt, B. (2016). Genome reduction uncovers a large dispensable genome and adaptive role for copy number variation in asexually propagated Solanum tuberosum. *The Plant Cell*, *28*(2), 388–405.

Hartemink, A. E., & Sonneveld, M. P. W. (2013). Soil maps of the Netherlands. *Geoderma*, *204*, 1–9.

Hashida, S.-N., Uchiyama, T., Martin, C., Kishima, Y., Sano, Y., & Mikami, T. (2006). The temperature-dependent change in methylation of the Antirrhinum transposon Tam3 is controlled by the activity of its transposase. *The Plant Cell*, *18*(1), 104–118.

Hastings, P. J., Ira, G., & Lupski, J. R. (2009). A microhomology-mediated break-induced replication model for the origin of human copy number variation. *PLoS Genetics*, *5*(1), e1000327.

He, X., & Zhang, J. (2005). Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics*, *169*(2), 1157–1164.

He, Z., Ji, R., Havlickova, L., Wang, L., Li, Y., Lee, H. T., Song, J., Koh, C., Yang, J., & Zhang, M. (2021). Genome structural evolution in Brassica crops. *Nature Plants*, *7*(6), 757–765.

Heddad, M., Norén, H., Reiser, V., Dunaeva, M., Andersson, B., & Adamska, I. (2006). Differential expression and localization of early light-induced proteins in Arabidopsis. *Plant Physiology*, *142*(1), 75–87.

Hell, R., & Stephan, U. W. (2003). Iron uptake, trafficking and homeostasis in plants. *Planta*, *216*, 541–551.

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., & Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, *25*(15), 1965–1978.

Hirsch, C. N., Hirsch, C. D., Brohammer, A. B., Bowman, M. J., Soifer, I., Barad, O., Shem-Tov, D., Baruch, K., Lu, F., & Hernandez, A. G. (2016).

Draft assembly of elite inbred line PH207 provides insights into genomic and transcriptome diversity in maize. *The Plant Cell*, *28*(11), 2700–2714.

Ho, S. S., Urban, A. E., & Mills, R. E. (2020). Structural variation in the sequencing era. *Nature Reviews Genetics*, *21*(3), 171–189.

Horton, M. W., Hancock, A. M., Huang, Y. S., Toomajian, C., Atwell, S., Auton, A., Muliyati, N. W., Platt, A., Sperone, F. G., & Vilhjálmsson, B. J. (2012). Genome-wide patterns of genetic variation in worldwide Arabidopsis thaliana accessions from the RegMap panel. *Nature Genetics*, *44*(2), 212–216.

Hsu, C., Lo, C., & Lee, C. (2019). On the postglacial spread of human commensal Arabidopsis thaliana: journey to the East. *New Phytologist*, *222*(3), 1447–1457.

Huang, C.-H., Sun, R., Hu, Y., Zeng, L., Zhang, N., Cai, L., Zhang, Q., Koch, M. A., Al-Shehbaz, I., & Edger, P. P. (2016). Resolution of Brassicaceae phylogeny using nuclear genes uncovers nested radiations and supports convergent morphological evolution. *Molecular Biology and Evolution*, *33*(2), 394–412.

Huang, H., Ullah, F., Zhou, D.-X., Yi, M., & Zhao, Y. (2019). Mechanisms of ROS regulation of plant development and stress responses. *Frontiers in Plant Science*, *10*, 800.

Huang, X., & Han, B. (2014). Natural variations and genome-wide association studies in crop plants. *Annual Review of Plant Biology*, *65*, 531–551.

Huang, X., Zhao, Y., Wei, X., Li, C., Wang, A., Zhao, Q., Li, W., Guo, Y., Deng, L., & Zhu, C. (2012). Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. *Nature Genetics*, *44*(1), 32–39.

Huddleston, J., & Eichler, E. E. (2016). An incomplete understanding of human genetic variation. *Genetics*, *202*(4), 1251–1254.

Hull, R. M., Cruz, C., Jack, C. V, & Houseley, J. (2017). Environmental change drives accelerated adaptation through stimulated copy number variation. *PLoS Biology*, *15*(6), e2001333.

Hutin, C., Nussaume, L., Moise, N., Moya, I., Kloppstech, K., & Havaux, M. (2003). Early light-induced proteins protect Arabidopsis from photooxidative stress. *Proceedings of the National Academy of Sciences*, *100*(8), 4921–4926.

Iakovishina, D., Janoueix-Lerosey, I., Barillot, E., Regnier, M., & Boeva, V. (2016). SV-Bay: structural variant detection in cancer genomes using a Bayesian approach with correction for GC-content and read mappability. *Bioinformatics*, *32*(7), 984–992.

Imprialou, M., Kahles, A., Steffen, J. G., Osborne, E. J., Gan, X., Lempe, J., Bhomra, A., Belfield, E., Visscher, A., & Greenhalgh, R. (2017). Genomic rearrangements in Arabidopsis considered as quantitative traits. *Genetics*, *205*(4), 1425–1441.

Iqbal, M., & Ashraf, M. (2007). Seed preconditioning modulates growth, ionic relations, and photosynthetic capacity in adult plants of hexaploid wheat under salt stress. *Journal of Plant Nutrition*, *30*(3), 381–396.

Ito, H., Gaubert, H., Bucher, E., Mirouze, M., Vaillant, I., & Paszkowski, J. (2011). An siRNA pathway prevents transgenerational retrotransposition in plants subjected to stress. *Nature*, *472*(7341), 115–119.

Jako, C., Kumar, A., Wei, Y., Zou, J., Barton, D. L., Giblin, E. M., Covello, P. S., & Taylor, D. C. (2001). Seed-specific over-expression of an Arabidopsis cDNA encoding a diacylglycerol acyltransferase enhances seed oil content and seed weight. *Plant Physiology*, *126*(2), 861–874.

Jeffares, D. C., Jolly, C., Hoti, M., Speed, D., Shaw, L., Rallis, C., Balloux, F., Dessimoz, C., Bähler, J., & Sedlazeck, F. J. (2017). Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nature Communications*, *8*(1), 14061.

Jiang, C., Mithani, A., Belfield, E. J., Mott, R., Hurst, L. D., & Harberd, N. P. (2014). Environmentally responsive genome-wide accumulation of de novo Arabidopsis thaliana mutations and epimutations. *Genome Research*, *24*(11), 1821–1829.

Jiang, N., Bao, Z., Zhang, X., Eddy, S. R., & Wessler, S. R. (2004). Pack-MULE transposable elements mediate gene evolution in plants. *Nature*, *431*(7008), 569–573.

Jiao, W. B., & Schneeberger, K. (2017). The impact of third generation genomic technologies on plant genome assembly. In *Current Opinion in Plant Biology* (Vol. 36, pp. 64–70). https://doi.org/10.1016/j.pbi.2017.02.002

Jiao, W.-B., & Schneeberger, K. (2020). Chromosome-level assemblies of multiple Arabidopsis genomes reveal hotspots of rearrangements with altered evolutionary dynamics. *Nature Communications*, *11*(1), 989.

Johannes, F., & Schmitz, R. J. (2019). Spontaneous epimutations in plants. *New Phytologist*, *221*(3), 1253–1259.

Johanson, U., West, J., Lister, C., Michaels, S., Amasino, R., & Dean, C. (2000). Molecular analysis of FRIGIDA, a major determinant of natural variation in Arabidopsis flowering time. *Science*, *290*(5490), 344–347.

Johnston, J. S., Pepper, A. E., Hall, A. E., Chen, Z. J., Hodnett, G., Drabek, J., Lopez, R., & Price, H. J. (2005). Evolution of genome size in Brassicaceae. *Annals of Botany*, *95*(1), 229–235.

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., & Nuka, G. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics*, *30*(9), 1236–1240.

Joseph, J. T., Poolakkalody, N. J., & Shah, J. M. (2018). Plant reference genes for development and stress response studies. *Journal of Biosciences*, *43*, 173–187.

Jugulam, M., Niehues, K., Godar, A. S., Koo, D.-H., Danilova, T., Friebe, B., Sehgal, S., Varanasi, V. K., Wiersma, A., & Westra, P. (2014). Tandem amplification of a chromosomal segment harboring 5-enolpyruvylshikimate-3-phosphate synthase locus confers glyphosate resistance in Kochia scoparia. *Plant Physiology*, *166*(3), 1200–1207.

Juretic, N., Hoen, D. R., Huynh, M. L., Harrison, P. M., & Bureau, T. E. (2005). The evolutionary fate of MULE-mediated duplications of host gene fragments in rice. *Genome Research*, *15*(9), 1292–1297.

Kadota, Y., Shirasu, K., & Zipfel, C. (2015). Regulation of the NADPH oxidase RBOHD during plant immunity. *Plant and Cell Physiology*, *56*(8), 1472–1480.

Kaiser, E., Correa Galvis, V., & Armbruster, U. (2019). Efficient photosynthesis in dynamic light environments: a chloroplast's perspective. *Biochemical Journal*, *476*(19), 2725–2741.

Kaiser, E., Morales, A., & Harbinson, J. (2018). Fluctuating light takes crop photosynthesis on a rollercoaster ride. *Plant Physiology*, *176*(2), 977–989.

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.

Katche, E., Quezada-Martinez, D., Katche, E. I., Vasquez-Teuber, P., & Mason, A. S. (2019). Interspecific hybridization for Brassica crop improvement. *Crop Breeding, Genetics and Genomics*, *1*(1).

Kato, M. C., Hikosaka, K., Hirotsu, N., Makino, A., & Hirose, T. (2003). The excess light energy that is neither utilized in photosynthesis nor dissipated by photoprotective mechanisms determines the rate of photoinactivation in photosystem II. *Plant and Cell Physiology*, *44*(3), 318–325.

Kaur, G., Sharma, A., Guruprasad, K., & Pati, P. K. (2014). Versatile roles of plant NADPH oxidases and emerging concepts. *Biotechnology Advances*, *32*(3), 551–563.

Kawecki, T. J., & Ebert, D. (2004). Conceptual issues in local adaptation. *Ecology Letters*, *7*(12), 1225–1241.

Kawecki, T. J., Lenski, R. E., Ebert, D., Hollis, B., Olivieri, I., & Whitlock, M. C. (2012). Experimental evolution. *Trends in Ecology & Evolution*, *27*(10), 547–560.

Kent, T. V, Uzunović, J., & Wright, S. I. (2017). Coevolution between transposable elements and recombination. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1736), 20160458.

Kidd, J. M., Sampas, N., Antonacci, F., Graves, T., Fulton, R., Hayden, H. S., Alkan, C., Malig, M., Ventura, M., & Giannuzzi, G. (2010). Characterization of missing human genome sequences and copy-number polymorphic insertions. *Nature Methods*, *7*(5), 365–371.

Kiełbasa, S. M., Wan, R., Sato, K., Horton, P., & Frith, M. C. (2011). Adaptive seeds tame genomic sequence comparison. *Genome Research*, *21*(3), 487–493.

Kim, H., Choi, S. R., Bae, J., Hong, C. P., Lee, S. Y., Hossain, M. J., Van Nguyen, D., Jin, M., Park, B.-S., & Bang, J.-W. (2009). Sequenced BAC anchored reference genetic map that reconciles the ten individual chromosomes of Brassica rapa. *BMC Genomics*, *10*(1), 1–15.

Koester, R. P., Nohl, B. M., Diers, B. W., & Ainsworth, E. A. (2016). Has photosynthetic capacity increased with 80 years of soybean breeding? An examination of historical soybean cultivars. *Plant, Cell & Environment*, *39*(5), 1058–1067.

Kole, C. (2011). *Wild crop relatives: genomic and breeding resources*. Springer.

Koornneef, M., Blankestijn-de Vries, H., Hanhart, C., Soppe, W., & Peeters, T. (1994). The phenotype of some late-flowering mutants is enhanced by a locus on chromosome 5 that is not effective in the Landsberg erecta wild-type. *The Plant Journal*, *6*(6), 911–919.

Koornneef, M., & Meinke, D. (2010). The development of Arabidopsis as a model plant. *The Plant Journal*, *61*(6), 909–921.

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*, *27*(5), 722–736.

Korte, A., & Farlow, A. (2013). The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods*, *9*(1), 1–9.

Krämer, U. (2010). Metal hyperaccumulation in plants. *Annual Review of Plant Biology*, *61*, 517–534.

Kromdijk, J., Głowacka, K., Leonelli, L., Gabilly, S. T., Iwai, M., Niyogi, K. K., & Long, S. P. (2016). Improving photosynthesis and crop productivity by accelerating recovery from photoprotection. *Science*, *354*(6314), 857–861.

Krueger, F., & Andrews, S. R. (2011). Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*, *27*(11), 1571–1572.

Kwak, J. M., Mori, I. C., Pei, Z.-M., Leonhardt, N., Torres, M. A., Dangl, J. L., Bloom, R. E., Bodde, S., Jones, J. D. G., & Schroeder, J. I. (2003). NADPH oxidase AtrbohD and AtrbohF genes function in ROS-dependent ABA signaling in Arabidopsis. *The EMBO Journal*, *22*(11), 2623–2633.

Lai, J., Li, R., Xu, X., Jin, W., Xu, M., Zhao, H., Xiang, Z., Song, W., Ying, K., & Zhang, M. (2010). Genome-wide patterns of genetic variation among elite maize inbred lines. *Nature Genetics*, *42*(11), 1027–1030.

Lai, J., Li, Y., Messing, J., & Dooner, H. K. (2005). Gene movement by Helitron transposons contributes to the haplotype variability of maize. *Proceedings of the National Academy of Sciences*, *102*(25), 9068–9073.

Lämke, J., & Bäurle, I. (2017). Epigenetic and chromatin-based mechanisms in environmental stress adaptation and stress memory in plants. *Genome Biology*, *18*(1), 1–11.

Lanciano, S., & Mirouze, M. (2018). Transposable elements: all mobile, all different, some stress responsive, some adaptive? *Current Opinion in Genetics & Development*, *49*, 106–114.

Lauer, S., & Gresham, D. (2019). An evolving view of copy number variants. *Current Genetics*, *65*(6), 1287–1295.

Law, J. A., & Jacobsen, S. E. (2010). Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nature Reviews Genetics*, *11*(3), 204–220.

Lawson, T., Kramer, D. M., & Raines, C. A. (2012). Improving yield by exploiting mechanisms underlying natural variation of photosynthesis. *Current Opinion in Biotechnology*, *23*(2), 215–220.

Layer, R. M., Chiang, C., Quinlan, A. R., & Hall, I. M. (2014). LUMPY: a probabilistic framework for structural variant discovery. *Genome Biology*, *15*(6), 1–19.

Le, C. T. T., Brumbarova, T., Ivanov, R., Stoof, C., Weber, E., Mohrbacher, J., Fink-Straube, C., & Bauer, P. (2016). Zinc finger of Arabidopsis thaliana12 (ZAT12) interacts with FER-like iron deficiency-induced transcription factor (FIT) linking iron deficiency and oxidative stress responses. *Plant Physiology*, *170*(1), 540–557.

Le Corre, V., Roux, F., & Reboud, X. (2002). DNA polymorphism at the FRIGIDA gene in Arabidopsis thaliana: extensive nonsynonymous variation is consistent with local selection for flowering time. *Molecular Biology and Evolution*, *19*(8), 1261–1271.

Leakey, A. D. B., Uribelarrea, M., Ainsworth, E. A., Naidu, S. L., Rogers, A., Ort, D. R., & Long, S. P. (2006). Photosynthesis, productivity, and yield of maize are not affected by open-air elevation of CO2 concentration in the absence of drought. *Plant Physiology*, *140*(2), 779–790.

Lee, C.-R., Svardal, H., Farlow, A., Exposito-Alonso, M., Ding, W., Novikova, P., Alonso-Blanco, C., Weigel, D., & Nordborg, M. (2017). On the post-glacial spread of human commensal Arabidopsis thaliana. *Nature Communications*, *8*(1), 14458.

Lee, J. H., & Jeon, J. T. (2009). Methods to detect and analyze copy number variations at the genome-wide and locus-specific levels. *Cytogenetic and Genome Research*, *123*(1–4), 333–342.

Lee, P. L. M., Patel, R. M., Conlan, R. S., Wainwright, S. J., & Hipkin, C. R. (2004). Comparison of genetic diversities in native and alien populations of hoary mustard (Hirschfeldia incana [L.] Lagreze-Fossat). *International Journal of Plant Sciences*, *165*(5), 833–843.

Lee, S., Jeon, J., Boernke, F., Voll, L., Cho, J., Goh, C., Jeong, S., Park, Y., Kim, S. J., & Choi, S. (2008). Loss of cytosolic fructose-1, 6-bisphosphatase limits photosynthetic sucrose synthesis and causes severe growth retardations in rice (Oryza sativa). *Plant, Cell & Environment*, *31*(12), 1851–1863.

Lee, S., Joung, Y. H., Kim, J.-K., Do Choi, Y., & Jang, G. (2019). An isoform of the plastid RNA polymerase-associated protein FSD3 negatively regulates chloroplast development. *BMC Plant Biology*, *19*(1), 1–12.

Lee, T. G., Diers, B. W., & Hudson, M. E. (2016). An efficient method for measuring copy number variation applied to improvement of nematode resistance in soybean. *The Plant Journal*, *88*(1), 143–153.

Leister, D. (2019). Thawing out frozen metabolic accidents. *BMC Biology*, *17*, 1–11.

Lenski, R. E., Rose, M. R., Simpson, S. C., & Tadler, S. C. (1991). Long-term experimental evolution in Escherichia coli. I. Adaptation and divergence during 2,000 generations. *The American Naturalist*, *138*(6), 1315–1341.

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.

Li, G.-M. (2008). Mechanisms and functions of DNA mismatch repair. *Cell Research*, *18*(1), 85–98.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv Preprint ArXiv:1303.3997*.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & Subgroup, 1000 Genome Project Data Processing. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079.

Li, Y., Huang, Y., Bergelson, J., Nordborg, M., & Borevitz, J. O. (2010). Association mapping of local climate-sensitive quantitative trait loci in Arabidopsis thaliana. *Proceedings of the National Academy of Sciences*, *107*(49), 21199–21204.

Li, Y., Liu, B., Zhang, J., Kong, F., Zhang, L., Meng, H., Li, W., Rochaix, J.-D., Li, D., & Peng, L. (2019). OHP1, OHP2, and HCF244 form a transient functional complex with the photosystem II reaction center. *Plant Physiology*, *179*(1), 195–208.

Li, Y., Zhou, G., Ma, J., Jiang, W., Jin, L., Zhang, Z., Guo, Y., Zhang, J., Sui, Y., & Zheng, L. (2014). De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nature Biotechnology*, *32*(10), 1045–1052.

Li, Z., Defoort, J., Tasdighian, S., Maere, S., Van de Peer, Y., & De Smet, R. (2016). Gene duplicability of core genes is highly consistent across all angiosperms. *The Plant Cell*, *28*(2), 326–344.

Lian, Q., Hüttel, B., Walkemeier, B., Mayjonade, B., Lopez-Roques, C., Gil, L., Roux, F., Schneeberger, K., & Mercier, R. (2023). *A pan-genome of 72 Arabidopsis thaliana accessions reveals a conserved genome structure throughout the global species range*.

Lin, K., Smit, S., Bonnema, G., Sanchez-Perez, G., & de Ridder, D. (2015). Making the difference: integrating structural variation detection tools. *Briefings in Bioinformatics*, *16*(5), 852–864.

Lin, K., Zhang, N., Severing, E. I., Nijveen, H., Cheng, F., Visser, R. G. F., Wang, X., de Ridder, D., & Bonnema, G. (2014). Beyond genomic variation-comparison and functional annotation of three Brassica rapa genomes: a turnip, a rapid cycling and a Chinese cabbage. *Bmc Genomics*, *15*, 1–17.

Liu, H., Able, A. J., & Able, J. A. (2022). Priming crops for the future: rewiring stress memory. *Trends in Plant Science*, *27*(7), 699–716.

Liu, H., & Yan, J. (2019). Crop genome-wide association study: a harvest of biological relevance. *The Plant Journal*, *97*(1), 8–18.

Liu, S., Liu, Y., Yang, X., Tong, C., Edwards, D., Parkin, I. A. P., Zhao, M., Ma, J., Yu, J., & Huang, S. (2014). The Brassica oleracea genome reveals the asymmetrical evolution of polyploid genomes. *Nature Communications*, *5*(1), 3930.

Liu, Y., Wei, W., Ma, K., Li, J., Liang, Y., & Darmency, H. (2013). Consequences of gene flow between oilseed rape (Brassica napus) and its relatives. *Plant Science*, *211*, 42–51.

Livak, K. J., & Schmittgen, T. D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2− ΔΔCT method. *Methods*, *25*(4), 402–408.

Long, Q., Rabanal, F. A., Meng, D., Huber, C. D., Farlow, A., Platzer, A., Zhang, Q., Vilhjálmsson, B. J., Korte, A., & Nizhynska, V. (2013). Massive genomic variation and strong selection in Arabidopsis thaliana lines from Sweden. *Nature Genetics*, *45*(8), 884–890.

Long, S. P., Marshall-Colon, A., & Zhu, X.-G. (2015). Meeting the global food demand of the future by engineering crop photosynthesis and yield potential. *Cell*, *161*(1), 56–66.

Long, S. P., ZHU, X., Naidu, S. L., & Ort, D. R. (2006). Can improvement in photosynthesis increase crop yields? *Plant, Cell & Environment*, *29*(3), 315–330.

Løvdal, T., & Lillo, C. (2009). Reference gene selection for quantitative real-time PCR normalization in tomato subjected to nitrogen, cold, and light stress. *Analytical Biochemistry*, *387*(2), 238–242.

Lu, F., Romay, M. C., Glaubitz, J. C., Bradbury, P. J., Elshire, R. J., Wang, T., Li, Y., Li, Y., Semagn, K., & Zhang, X. (2015). High-resolution genetic mapping of maize pan-genome sequence anchors. *Nature Communications*, *6*(1), 1–8.

Lu, P., Han, X., Qi, J., Yang, J., Wijeratne, A. J., Li, T., & Ma, H. (2012). Analysis of Arabidopsis genome-wide variations before and after meiosis and meiotic recombination by resequencing Landsberg erecta and all four products of a single meiosis. *Genome Research*, *22*(3), 508–518.

Lu, Z., Cui, J., Wang, L., Teng, N., Zhang, S., Lam, H.-M., Zhu, Y., Xiao, S., Ke, W., & Lin, J. (2021). Genome-wide DNA mutations in Arabidopsis plants after multigenerational exposure to high temperatures. *Genome Biology*, *22*(1), 160.

Lucht, J. M., Mauch-Mani, B., Steiner, H.-Y., Metraux, J.-P., Ryals, J., & Hohn, B. (2002). Pathogen stress increases somatic recombination frequency in Arabidopsis. *Nature Genetics*, *30*(3), 311–314.

Luo, Y., Dong, X., Yu, T., Shi, X., Li, Z., Yang, W., Widmer, A., & Karrenberg, S. (2015). A single nucleotide deletion in gibberellin20-oxidase1 causes alpine dwarfism in Arabidopsis. *Plant Physiology*, *168*(3), 930–937.

Lye, Z. N., & Purugganan, M. D. (2019). Copy number variation in domestication. *Trends in Plant Science*, *24*(4), 352–365.

Lysak, M. A., Koch, M. A., Pecinka, A., & Schubert, I. (2005). Chromosome triplication found across the tribe Brassiceae. *Genome Research*, *15*(4), 516–525.

Mackay, T. F. C. (2001). The genetic architecture of quantitative traits. *Annual Review of Genetics*, *35*(1), 303–339.

Madlung, A., & Comai, L. (2004). The effect of stress on genome regulation and structure. *Annals of Botany*, *94*(4), 481–495.

Maiwald, D., Dietzmann, A., Jahns, P., Pesaresi, P., Joliot, P., Joliot, A., Levin, J. Z., Salamini, F., & Leister, D. (2003). Knock-out of the genes coding for the Rieske protein and the ATP-synthase δ-subunit of Arabidopsis. Effects on photosynthesis, thylakoid protein composition, and nuclear chloroplast gene expression. *Plant Physiology*, *133*(1), 191–202.

Maksymiec, W., Wianowska, D., Dawidowicz, A. L., Radkiewicz, S., Mardarowicz, M., & Krupa, Z. (2005). The level of jasmonic acid in Arabidopsis thaliana and Phaseolus coccineus plants under heavy metal stress. *Journal of Plant Physiology*, *162*(12), 1338–1346.

Mandakova, T., Li, Z., Barker, M. S., & Lysak, M. A. (2017). Diverse genome organization following 13 independent mesopolyploid events in Brassicaceae contrasts with convergent patterns of gene retention. *The Plant Journal*, *91*(1), 3–21.

Mandáková, T., & Lysak, M. A. (2018). Post-polyploid diploidization and diversification through dysploid changes. *Current Opinion in Plant Biology*, *42*, 55–65.

Mangeon, A., Pardal, R., Menezes-Salgueiro, A. D., Duarte, G. L., de Seixas, R., Cruz, F. P., Cardeal, V., Magioli, C., Ricachenevsky, F. K., & Margis, R. (2016). AtGRP3 is implicated in root size and aluminum response pathways in Arabidopsis. *PLoS One*, *11*(3), e0150583.

Mansfeld, B. N., & Grumet, R. (2018). QTLseqr: An R package for bulk segregant analysis with next-generation sequencing. *The Plant Genome*, *11*(2), 180006.

Mansisidor, A., Molinar, T., Srivastava, P., Dartis, D. D., Delgado, A. P., Blitzblau, H. G., Klein, H., & Hochwagen, A. (2018). Genomic copy-number loss is rescued by self-limiting production of DNA circles. *Molecular Cell*, *72*(3), 583–593.

Marçais, G., & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, *27*(6), 764–770.

Maron, L. G., Guimarães, C. T., Kirst, M., Albert, P. S., Birchler, J. A., Bradbury, P. J., Buckler, E. S., Coluccio, A. E., Danilova, T. V, & Kudrna, D. (2013). Aluminum tolerance in maize is associated with higher MATE1 gene copy number. *Proceedings of the National Academy of Sciences*, *110*(13), 5241–5246.

Marroni, F., Pinosio, S., & Morgante, M. (2014). Structural variation and genome complexity: is dispensable really dispensable? *Current Opinion in Plant Biology*, *18*, 31–36.

Marshall, B., & Biscoe, P. V. (1980). A model for C3 leaves describing the dependence of net photosynthesis on irradiance. *Journal of Experimental Botany*, *31*(1), 29–39.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. Journal*, *17*(1), 10–12.

Martínez-Castilla, L. P., & Alvarez-Buylla, E. R. (2003). Adaptive evolution in the Arabidopsis MADS-box gene family inferred from its complete resolved phylogeny. *Proceedings of the National Academy of Sciences*, *100*(23), 13407–13412.

Martinis, J., Glauser, G., Valimareanu, S., & Kessler, F. (2013). A chloroplast ABC1-like kinase regulates vitamin E metabolism in Arabidopsis. *Plant Physiology*, *162*(2), 652–662.

Maruta, T., Inoue, T., Tamoi, M., Yabuta, Y., Yoshimura, K., Ishikawa, T., & Shigeoka, S. (2011). Arabidopsis NADPH oxidases, AtrbohD and AtrbohF, are essential for jasmonic acid-induced expression of genes regulated by MYC2 transcription factor. *Plant Science*, *180*(4), 655–660.

Matsuba, C., Ostrow, D. G., Salomon, M. P., Tolani, A., & Baer, C. F. (2013). Temperature, stress and spontaneous mutation in Caenorhabditis briggsae and Caenorhabditis elegans. *Biology Letters*, *9*(1), 20120334.

Maxwel, K., & Johnson, G. N. (2000). Chlorophyll fluorescence—a practical guide. *Journal of Experimental Botany*, *51*(345), 659–668.

McAusland, L., Vialet-Chabrand, S., Jauregui, I., Burridge, A., Hubbart-Edwards, S., Fryer, M. J., King, I. P., King, J., Pyke, K., & Edwards, K. J. (2020). Variation in key leaf photosynthetic traits across wheat wild relatives is accession dependent not species dependent. *New Phytologist*, *228*(6), 1767–1780.

McCarroll, S. A., & Altshuler, D. M. (2007). Copy-number variation and association studies of human disease. *Nature Genetics*, *39*(Suppl 7), S37–S42.

McHale, L. K., Haun, W. J., Xu, W. W., Bhaskar, P. B., Anderson, J. E., Hyten, D. L., Gerhardt, D. J., Jeddeloh, J. A., & Stupar, R. M. (2012). Structural variants in the soybean genome localize to clusters of biotic stress-response genes. *Plant Physiology*, *159*(4), 1295–1308.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., & Daly, M. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303.

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., Flicek, P., & Cunningham, F. (2016). The ensembl variant effect predictor. *Genome Biology*, *17*(1), 1–14.

Mertens, D., Boege, K., Kessler, A., Koricheva, J., Thaler, J. S., Whiteman, N. K., & Poelman, E. H. (2021). Predictability of biotic stress structures plant defence evolution. *Trends in Ecology & Evolution*, *36*(5), 444–456.

Mertens, D., Fernández de Bobadilla, M., Rusman, Q., Bloem, J., Douma, J. C., & Poelman, E. H. (2021). Plant defence to sequential attack is adapted to prevalent herbivores. *Nature Plants*, *7*(10), 1347–1353.

Mhamdi, A., & Van Breusegem, F. (2018). Reactive oxygen species in plant development. *Development*, *145*(15), dev164376.

Migicovsky, Z., & Kovalchuk, I. (2013). Changes to DNA methylation and homologous recombination frequency in the progeny of stressed plants. *Biochemistry and Cell Biology*, *91*(1), 1–5.

Miller, G., Schlauch, K., Tam, R., Cortes, D., Torres, M. A., Shulaev, V., Dangl, J. L., & Mittler, R. (2009). The plant NADPH oxidase RBOHD mediates rapid systemic signaling in response to diverse stimuli. *Science Signaling*, *2*(84), ra45–ra45.

Mira, S., Veiga-Barbosa, L., & Pérez-García, F. (2019). Seed dormancy and longevity variability of Hirschfeldia incana L. during storage. *Seed Science Research*, *29*(2), 97–103.

Mitchell, A. L., Attwood, T. K., Babbitt, P. C., Blum, M., Bork, P., Bridge, A., Brown, S. D., Chang, H. Y., El-Gebali, S., Fraser, M. I., Gough, J., Haft, D. R., Huang, H., Letunic, I., Lopez, R., Luciani, A., Madeira, F., Marchler-Bauer, A., Mi, H., … Finn, R. D. (2019). InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Research*, *47*(D1), D351–D360. https://doi.org/10.1093/NAR/GKY1100

Mittler, R. (2006). Abiotic stress, the field environment and stress combination. *Trends in Plant Science*, *11*(1), 15–19.

Mladenov, V., Fotopoulos, V., Kaiserli, E., Karalija, E., Maury, S., Baranek, M., Segal, N., Testillano, P. S., Vassileva, V., & Pinto, G. (2021). Deciphering the epigenetic alphabet involved in transgenerational stress memory in crops. *International Journal of Molecular Sciences*, *22*(13), 7118.

Mohanta, T. K., Bashir, T., Hashem, A., Abd_Allah, E. F., Khan, A. L., & Al-Harrasi, A. S. (2018). Early events in plant abiotic stress signaling: interplay between calcium, reactive oxygen species and phytohormones. *Journal of Plant Growth Regulation*, *37*, 1033–1049.

Molin, W. T., Wright, A. A., Lawton-Rauh, A., & Saski, C. A. (2017). The unique genomic landscape surrounding the EPSPS gene in glyphosate resistant Amaranthus palmeri: a repetitive path to resistance. *Bmc Genomics*, *18*, 1–16.

Monnahan, P., Kolář, F., Baduel, P., Sailer, C., Koch, J., Horvath, R., Laenen, B., Schmickl, R., Paajanen, P., & Šrámková, G. (2019). Pervasive population genomic consequences of genome duplication in Arabidopsis arenosa. *Nature Ecology & Evolution*, *3*(3), 457–468.

Monneveux, P., Pastenes, C., & Reynolds, M. P. (2003). Limitations to photosynthesis under light and heat stress in three high-yielding wheat genotypes. *Journal of Plant Physiology*, *160*(6), 657–666.

Monteith, J. L. (1977). Climate and the efficiency of crop production in Britain. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, *281*(980), 277–294.

Morgante, M., De Paoli, E., & Radovic, S. (2007). Transposable elements and the plant pan-genomes. *Current Opinion in Plant Biology*, *10*(2), 149–155.

Munekage, Y., Hashimoto, M., Miyake, C., Tomizawa, K.-I., Endo, T., Tasaka, M., & Shikanai, T. (2004). Cyclic electron flow around photosystem I is essential for photosynthesis. *Nature*, *429*(6991), 579–582.

Muñoz-Amatriaín, M., Eichten, S. R., Wicker, T., Richmond, T. A., Mascher, M., Steuernagel, B., Scholz, U., Ariyadasa, R., Spannagl, M., & Nussbaumer, T. (2013). Distribution, functional impact, and origin mechanisms of copy number variation in the barley genome. *Genome Biology*, *14*, 1–17.

Murchie, E. H., Chen, Y., Hubbart, S., Peng, S., & Horton, P. (1999). Interactions between senescence and leaf orientation determine in situ patterns of photosynthesis and photoinhibition in field-grown rice. *Plant Physiology*, *119*(2), 553–564.

Myouga, F., Hosoda, C., Umezawa, T., Iizumi, H., Kuromori, T., Motohashi, R., Shono, Y., Nagata, N., Ikeuchi, M., & Shinozaki, K. (2008). A heterocomplex of iron superoxide dismutases defends chloroplast nucleoids against oxidative stress and is essential for chloroplast development in Arabidopsis. *The Plant Cell*, *20*(11), 3148–3162.

Nam, J., Kim, J., Lee, S., An, G., Ma, H., & Nei, M. (2004). Type I MADS-box genes have experienced faster birth-and-death evolution than type II MADS-box genes in angiosperms. *Proceedings of the National Academy of Sciences*, *101*(7), 1910–1915.

Nitcher, R., Distelfeld, A., Tan, C., Yan, L., & Dubcovsky, J. (2013). Increased copy number at the HvFT1 locus is associated with accelerated flowering time in barley. *Molecular Genetics and Genomics*, *288*, 261–275.

Noir, S., Bömer, M., Takahashi, N., Ishida, T., Tsui, T.-L., Balbi, V., Shanahan, H., Sugimoto, K., & Devoto, A. (2013). Jasmonate controls leaf growth by repressing cell proliferation and the onset of endoreduplication while maintaining a potential stand-by mode. *Plant Physiology*, *161*(4), 1930–1951.

Norén, H., Svensson, P., Stegmark, R., Funk, C., Adamska, I., & Andersson, B. (2003). Expression of the early light-induced protein but not the PsbS protein is influenced by low temperature and depends on the developmental stage of the plant in field-grown pea cultivars. *Plant, Cell & Environment*, *26*(2), 245–253.

Ó Lochlainn, S., Bowen, H. C., Fray, R. G., Hammond, J. P., King, G. J., White, P. J., Graham, N. S., & Broadley, M. R. (2011). Tandem quadruplication of HMA4 in the zinc (Zn) and cadmium (Cd) hyperaccumulator Noccaea caerulescens. *PloS One*, *6*(3), e17814.

Oh, D.-H., Dassanayake, M., Bohnert, H. J., & Cheeseman, J. M. (2013). Life at the extreme: lessons from the genome. *Genome Biology*, *13*(3), 1–9.

Ohno, S. (1970). *Evolution by gene duplication*. Springer Science & Business Media.

Orr, H. A. (1998). The population genetics of adaptation: the distribution of factors fixed during adaptive evolution. *Evolution*, *52*(4), 935–949.

Ort, D. R., Merchant, S. S., Alric, J., Barkan, A., Blankenship, R. E., Bock, R., Croce, R., Hanson, M. R., Hibberd, J. M., & Long, S. P. (2015). Redesigning photosynthesis to sustainably meet global food and bioenergy demand. *Proceedings of the National Academy of Sciences*, *112*(28), 8529–8536.

Ortiz, D., Hu, J., & Salas Fernandez, M. G. (2017). Genetic architecture of photosynthesis in Sorghum bicolor under non-stress and cold stress conditions. *Journal of Experimental Botany*, *68*(16), 4545–4557.

Ossowski, S., Schneeberger, K., Lucas-Lledó, J. I., Warthmann, N., Clark, R. M., Shaw, R. G., Weigel, D., & Lynch, M. (2010). The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana. *Science*, *327*(5961), 92–94.

Ou, S., & Jiang, N. (2018). LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiology*, *176*(2), 1410–1422.

Oud, B., Guadalupe-Medina, V., Nijkamp, J. F., de Ridder, D., Pronk, J. T., van Maris, A. J. A., & Daran, J.-M. (2013). Genome duplication and mutations in ACE2 cause multicellular, fast-sedimenting phenotypes in evolved Saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences*, *110*(45), E4223–E4231.

Papp, B., Pál, C., & Hurst, L. D. (2003). Dosage sensitivity and the evolution of gene families in yeast. *Nature*, *424*(6945), 194–197.

Paradis, E., & Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, *35*(3), 526–528.

Parenicova, L., de Folter, S., Kieffer, M., Horner, D. S., Favalli, C., Busscher, J., Cook, H. E., Ingram, R. M., Kater, M. M., & Davies, B. (2003). Molecular and phylogenetic analyses of the complete MADS-box transcription factor family in Arabidopsis: new openings to the MADS world. *The Plant Cell*, *15*(7), 1538–1551.

Paritosh, K., Pradhan, A. K., & Pental, D. (2020). A highly contiguous genome assembly of Brassica nigra (BB) and revised nomenclature for the pseudochromosomes. *BMC Genomics*, *21*(1), 1–12.

Patterson, E. L., Pettinga, D. J., Ravet, K., Neve, P., & Gaines, T. A. (2018). Glyphosate resistance and EPSPS gene duplication: convergent evolution in multiple plant species. *Journal of Heredity*, *109*(2), 117–125.

Pauli, D., Ziegler, G., Ren, M., Jenks, M. A., Hunsaker, D. J., Zhang, M., Baxter, I., & Gore, M. A. (2018). Multivariate analysis of the cotton seed ionome reveals a shared genetic architecture. *G3: Genes, Genomes, Genetics*, *8*(4), 1147–1160.

Pedersen, B. S., & Quinlan, A. R. (2018). Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics*, *34*(5), 867–868.

Pedersen, B. S., & Quinlan, A. R. (2019). Duphold: scalable, depth-based annotation and curation of high-confidence structural variant calls. *Gigascience*, *8*(4), giz040.

Peng, J.-S., Guan, Y.-H., Lin, X.-J., Xu, X.-J., Xiao, L., Wang, H.-H., & Meng, S. (2021). Comparative understanding of metal hyperaccumulation in plants: a mini-review. *Environmental Geochemistry and Health*, *43*, 1599–1607.

Peng, L., Fukao, Y., Fujiwara, M., Takami, T., & Shikanai, T. (2009). Efficient operation of NAD (P) H dehydrogenase requires supercomplex formation with photosystem I via minor LHCI in Arabidopsis. *The Plant Cell*, *21*(11), 3623–3640.

Pereira, A. (2016). Plant abiotic stress challenges from the changing environment. In *Frontiers in plant science* (Vol. 7, p. 1123). Frontiers Media SA.

Perumal, S., Koh, C. S., Jin, L., Buchwaldt, M., Higgins, E. E., Zheng, C., Sankoff, D., Robinson, S. J., Kagale, S., & Navabi, Z.-K. (2020). A high-contiguity Brassica nigra genome localizes active centromeres and defines the ancestral Brassica genome. *Nature Plants*, *6*(8), 929–941.

Pico, F. X., Méndez-Vigo, B., Martínez-Zapater, J. M., & Alonso-Blanco, C. (2008). Natural genetic variation of Arabidopsis thaliana is geographically structured in the Iberian Peninsula. *Genetics*, *180*(2), 1009–1021.

Pinosio, S., Giacomello, S., Faivre-Rampant, P., Taylor, G., Jorge, V., Le Paslier, M. C., Zaina, G., Bastien, C., Cattonaro, F., & Marroni, F. (2016). Characterization of the poplar pan-genome by genome-wide identification of structural variation. *Molecular Biology and Evolution*, *33*(10), 2706–2719.

Pleban, J. R., Mackay, D. S., Aston, T. L., Ewers, B. E., & Weinig, C. (2018). Phenotypic trait identification using a multimodel Bayesian method: a case study using photosynthesis in Brassica rapa genotypes. *Frontiers in Plant Science*, *9*, 448.

Powles, S. B. (1984). Photoinhibition of photosynthesis induced by visible light. *Annual Review of Plant Physiology*, *35*(1), 15–44.

Preuss, D., Rhee, S. Y., & Davis, R. W. (1994). Tetrad analysis possible in Arabidopsis with mutation of the QUARTET (QRT) genes. *Science*, *264*(5164), 1458–1460.

Pucker, B., Kleinbölting, N., & Weisshaar, B. (2021). Large scale genomic rearrangements in selected Arabidopsis thaliana T-DNA lines are caused by T-DNA insertion mutagenesis. *BMC Genomics*, *22*(1), 1–21.

Purugganan, M. D., Rounsley, S. D., Schmidt, R. J., & Yanofsky, M. F. (1995). Molecular evolution of flower development: diversification of the plant MADS-box regulatory gene family. *Genetics*, *140*(1), 345–356.

Qian, J., Zheng, M., Wang, L., Song, Y., Yan, J., & Hsu, Y. (2022). Arabidopsis mitochondrial single-stranded DNA-binding proteins SSB1 and SSB2 are essential regulators of mtDNA replication and homologous recombination. *Journal of Integrative Plant Biology*, *64*(10), 1952–1965.

Qiao, X., Li, Q., Yin, H., Qi, K., Li, L., Wang, R., Zhang, S., & Paterson, A. H. (2019). Gene duplication and evolution in recurring polyploidization–diploidization cycles in plants. *Genome Biology*, *20*(1), 1–23.

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842.

Quiroz, D., Lensink, M., Kliebenstein, D. J., & Monroe, J. G. (2023). Causes of Mutation Rate Variability in Plant Genomes. *Annual Review of Plant Biology*, *74*, 751–775.

R Core Team, R. (2013). *R: A language and environment for statistical computing*.

Ram, Y., & Hadany, L. (2014). Stress-induced mutagenesis and complex adaptation. *Proceedings of the Royal Society B: Biological Sciences*, *281*(1792), 20141025.

Rausch, T., Zichner, T., Schlattl, A., Stütz, A. M., Benes, V., & Korbel, J. O. (2012). DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, *28*(18), i333–i339.

Rellstab, C., Gugerli, F., Eckert, A. J., Hancock, A. M., & Holderegger, R. (2015). A practical guide to environmental association analysis in landscape genomics. *Molecular Ecology*, *24*(17), 4348–4370.

Ritz, K. R., Noor, M. A. F., & Singh, N. D. (2017). Variation in recombination rate: adaptive or not? *Trends in Genetics*, *33*(5), 364–374.

Rivero, R. M., Mittler, R., Blumwald, E., & Zandalinas, S. I. (2022). Developing climate_resilient crops: improving plant tolerance to stress combination. *The Plant Journal*, *109*(2), 373–389.

Rizzon, C., Ponger, L., & Gaut, B. S. (2006). Striking similarities in the genomic distribution of tandemly arrayed genes in Arabidopsis and rice. *PLoS Computational Biology*, *2*(9), e115.

Ruban, A. V. (2017). Crops on the fast track for light. *Nature*, *541*(7635), 36–37.

Sammons, R. D., & Gaines, T. A. (2014). Glyphosate resistance: state of knowledge. *Pest Management Science*, *70*(9), 1367–1377.

Sánchez-Yélamo, M. D. (2009). Relationships in the Diplotaxis–Erucastrum–Brassica complex (Brassicaceae) evaluated from isoenzymatic profiles of the accessions as a whole. Applications for characterisation of phytogenetic resources preserved ex situ. *Genetic Resources and Crop Evolution*, *56*, 1023–1036.

Sankar, T. S., Wastuwidyaningtyas, B. D., Dong, Y., Lewis, S. A., & Wang, J. D. (2016). The nature of mutations induced by replication–transcription collisions. *Nature*, *535*(7610), 178–181.

Sasaki, E., Köcher, T., Filiault, D. L., & Nordborg, M. (2021). Revisiting a GWAS peak in Arabidopsis thaliana reveals possible confounding by genetic heterogeneity. *Heredity*, *127*(3), 245–252.

Scherer, S. W., Lee, C., Birney, E., Altshuler, D. M., Eichler, E. E., Carter, N. P., Hurles, M. E., & Feuk, L. (2007). Challenges and standards in integrating surveys of structural variation. *Nature Genetics*, *39*(Suppl 7), S7–S15.

Schippers, J. H. M., Nunes-Nesi, A., Apetrei, R., Hille, J., Fernie, A. R., & Dijkwel, P. P. (2008). The Arabidopsis onset of leaf death5 mutation of quinolinate synthase affects nicotinamide adenine dinucleotide biosynthesis and causes early ageing. *The Plant Cell*, *20*(10), 2909–2925.

Schlichting, C. D. (1986). The evolution of phenotypic plasticity in plants. *Annual Review of Ecology and Systematics*, *17*(1), 667–693.

Schmid, M. W., Heichinger, C., Coman Schmid, D., Guthörl, D., Gagliardini, V., Bruggmann, R., Aluri, S., Aquino, C., Schmid, B., & Turnbull, L. A. (2018). Contribution of epigenetic variation to adaptation in Arabidopsis. *Nature Communications*, *9*(1), 4446.

Schmuths, H., Meister, A., Horres, R., & Bachmann, K. (2004). Genome size variation among accessions of Arabidopsis thaliana. *Annals of Botany*, *93*(3), 317–321.

Schoen, D. J., & Schultz, S. T. (2019). Somatic mutation and evolution in plants. *Annual Review of Ecology, Evolution, and Systematics*, *50*, 49–73.

Schranz, M. E., Lysak, M. A., & Mitchell-Olds, T. (2006). The ABC's of comparative genomics in the Brassicaceae: building blocks of crucifer genomes. *Trends in Plant Science*, *11*(11), 535–542.

Sedlazeck, F. J., Lee, H., Darby, C. A., & Schatz, M. C. (2018). Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nature Reviews Genetics*, *19*(6), 329–346.

Sedlazeck, F. J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., Von Haeseler, A., & Schatz, M. C. (2018). Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods*, *15*(6), 461–468.

Seren, Ü., Vilhjálmsson, B. J., Horton, M. W., Meng, D., Forai, P., Huang, Y. S., Long, Q., Segura, V., & Nordborg, M. (2012). GWAPP: a web application for genome-wide association mapping in Arabidopsis. *The Plant Cell*, *24*(12), 4793–4805.

Sharma, P., Jha, A. B., Dubey, R. S., & Pessarakli, M. (2012). Reactive oxygen species, oxidative damage, and antioxidative defense mechanism in plants under stressful conditions. *Journal of Botany*, *2012*.

Sheikhizadeh Anari, S., de Ridder, D., Schranz, M. E., & Smit, S. (2018). Efficient inference of homologs in large eukaryotic pan-proteomes. *BMC Bioinformatics*, *19*(1), 1–11.

Sheikhizadeh, S., Schranz, M. E., Akdel, M., de Ridder, D., & Smit, S. (2016). PanTools: representation, storage and exploration of pan-genomic data. *Bioinformatics*, *32*(17), i487–i493.

Shen, X., & Carlborg, Ö. (2013). Beware of risk for increased false positive rates in genome-wide association studies for phenotypic variability. *Frontiers in Genetics*, *4*, 93.

Shi, T., Bibby, T. S., Jiang, L., Irwin, A. J., & Falkowski, P. G. (2005). Protein interactions limit the rate of evolution of photosynthetic genes in cyanobacteria. *Molecular Biology and Evolution*, *22*(11), 2179–2189.

Sieber, A.-N., Longin, C. F. H., Leiser, W. L., & Würschum, T. (2016). Copy number variation of CBF-A14 at the Fr-A2 locus determines frost tolerance in winter durum wheat. *Theoretical and Applied Genetics*, *129*, 1087–1097.

Simkin, A. J., López-Calcagno, P. E., & Raines, C. A. (2019). Feeding the world: improving photosynthetic efficiency for sustainable crop production. *Journal of Experimental Botany*, *70*(4), 1119–1140.

Simonich, M. T., & Innes, R. W. (1995). A disease resistance gene in Arabidopsis with specificity for the avrPph3 gene of Pseudomonas syringae pv. phaseolicola. *Molecular Plant-Microbe Interactions: MPMI*, *8*(4), 637–640.

Sinclair, T. R., Rufty, T. W., & Lewis, R. S. (2019). Increasing photosynthesis: unlikely solution for world food problem. *Trends in Plant Science*, *24*(11), 1032–1039.

Slattery, R. A., Walker, B. J., Weber, A. P. M., & Ort, D. R. (2018). The impacts of fluctuating light on crop performance. *Plant Physiology*, *176*(2), 990–1003.

Sloan, D. B., Wu, Z., & Sharbrough, J. (2018). Correction of persistent errors in Arabidopsis reference mitochondrial genomes. *The Plant Cell*, *30*(3), 525–527.

Smit, A. F. A., Hubley, R., & Green, P. (2015). *RepeatMasker Open-4.0. 2013–2015*.

Sniegowski, P. D., Gerrish, P. J., & Lenski, R. E. (1997). Evolution of high mutation rates in experimental populations of E. coli. *Nature*, *387*(6634), 703–705.

Sobel, J. M., Chen, G. F., Watt, L. R., & Schemske, D. W. (2010). The biology of speciation. *Evolution*, *64*(2), 295–315.

Soler, R., Badenes-Pérez, F. R., Broekgaarden, C., Zheng, S., David, A., Boland, W., & Dicke, M. (2012). Plant-mediated facilitation between a lea-feeding and a phloem-feeding insect in a brassicaceous plant: from insect performance to gene transcription. *Functional Ecology*, *26*(1), 156–166.

Sollars, E. S. A., Harper, A. L., Kelly, L. J., Sambles, C. M., Ramirez-Gonzalez, R. H., Swarbreck, D., Kaithakottil, G., Cooper, E. D., Uauy, C., & Havlickova, L. (2017). Genome sequence and genetic diversity of European ash trees. *Nature*, *541*(7636), 212–216.

Somerville, C., & Koornneef, M. (2002). A fortunate choice: the history of Arabidopsis as a model plant. *Nature Reviews Genetics*, *3*(11), 883–889.

Spielmann, J., Ahmadi, H., Scheepers, M., Weber, M., Nitsche, S., Carnol, M., Bosman, B., Kroymann, J., Motte, P., & Clemens, S. (2020). The two copies of the zinc and cadmium ZIP6 transporter of Arabidopsis halleri have distinct effects on cadmium tolerance. *Plant, Cell & Environment*, *43*(9), 2143–2157.

Springer, N. M., Ying, K., Fu, Y., Ji, T., Yeh, C.-T., Jia, Y., Wu, W., Richmond, T., Kitzman, J., & Rosenbaum, H. (2009). Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. *PLoS Genetics*, *5*(11), e1000734.

Stadler, L. J. (1946). Spontaneous mutation at the R locus in maize. I. The aleurone-color and plant-color effects. *Genetics*, *31*(4), 377.

Startek, M., Szafranski, P., Gambin, T., Campbell, I. M., Hixson, P., Shaw, C. A., Stankiewicz, P., & Gambin, A. (2015). Genome-wide analyses of LINE–LINE-mediated nonallelic homologous recombination. *Nucleic Acids Research*, *43*(4), 2188–2198.

Steinrücken, H. C., & Amrhein, N. (1980). The herbicide glyphosate is a potent inhibitor of 5-enolpyruvylshikimic acid-3-phosphate synthase. *Biochemical and Biophysical Research Communications*, *94*(4), 1207–1212.

Stephenson, P., Baker, D., Girin, T., Perez, A., Amoah, S., King, G. J., & Østergaard, L. (2010). A rich TILLING resource for studying gene function in Brassica rapa. *BMC Plant Biology*, *10*(1), 1–10.

Stinchcombe, J. R., Weinig, C., Ungerer, M., Olsen, K. M., Mays, C., Halldorsdottir, S. S., Purugganan, M. D., & Schmitt, J. (2004). A latitudinal cline in flowering time in Arabidopsis thaliana modulated by the flowering time gene FRIGIDA. *Proceedings of the National Academy of Sciences*, *101*(13), 4712–4717.

Stinziano, J. R., Roback, C., Gamble, D., Murphy, B., Hudson, P., & Muir, C. D. (2020). Photosynthesis: tools for plant ecophysiology & modeling. *R Package Version*, *2*(1).
Strigens, A., Freitag, N. M., Gilbert, X., Grieder, C., Riedelsheimer, C., Schrag, T. A., Messmer, R., & Melchinger, A. E. (2013). Association mapping for chilling tolerance in elite flint and dent maize inbred lines evaluated in growth chamber and field experiments. *Plant, Cell & Environment*, *36*(10), 1871–1887.

Sun, H., Ding, J., Piednoël, M., & Schneeberger, K. (2018). findGSE: estimating genome size variation within human and Arabidopsis using k-mer frequencies. *Bioinformatics*, *34*(4), 550–557.

Suryawanshi, V., Talke, I. N., Weber, M., Eils, R., Brors, B., Clemens, S., & Krämer, U. (2016). Between-species differences in gene copy number are enriched among functions critical for adaptive evolution in Arabidopsis halleri. *BMC Genomics*, *17*(13), 43–64.

Sutton, T., Baumann, U., Hayes, J., Collins, N. C., Shi, B.-J., Schnurbusch, T., Hay, A., Mayo, G., Pallotta, M., & Tester, M. (2007). Boron-toxicity tolerance in barley arising from efflux transporter amplification. *Science*, *318*(5855), 1446–1449.

Suyama, M., Torrents, D., & Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research*, *34*(suppl_2), W609–W612.

Suzuki, N., Rivero, R. M., Shulaev, V., Blumwald, E., & Mittler, R. (2014). Abiotic and biotic stress combinations. *New Phytologist*, *203*(1), 32–43.

Swanson-Wagner, R. A., Eichten, S. R., Kumari, S., Tiffin, P., Stein, J. C., Ware, D., & Springer, N. M. (2010). Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. *Genome Research*, *20*(12), 1689–1699.

Tabas-Madrid, D., Méndez-Vigo, B., Arteaga, N., Marcer, A., Pascual-Montano, A., Weigel, D., Xavier Pico, F., & Alonso-Blanco, C. (2018). Genome-wide signatures of flowering adaptation to climate temperature: Regional analyses in a highly diverse native range of Arabidopsis thaliana. *Plant, Cell & Environment*, *41*(8), 1806–1820.

Tang, H., Wang, X., Bowers, J. E., Ming, R., Alam, M., & Paterson, A. H. (2008). Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Research*, *18*(12), 1944–1954.

Tasdighian, S., Van Bel, M., Li, Z., Van de Peer, Y., Carretero-Paulet, L., & Maere, S. (2017). Reciprocally retained genes in the angiosperm lineage show the hallmarks of dosage balance sensitivity. *The Plant Cell*, *29*(11), 2766–2785.

Taylor, S. H., Orr, D. J., Carmo-Silva, E., & Long, S. P. (2020). During photosynthetic induction, biochemical and stomatal limitations differ between Brassica crops. *Plant, Cell & Environment*, *43*(11), 2623–2636.

Terés, J., Busoms, S., Perez Martín, L., Luís-Villarroya, A., Flis, P., Álvarez-Fernández, A., Tolrà, R., Salt, D. E., & Poschenrieder, C. (2019). Soil carbonate drives local adaptation in Arabidopsis thaliana. *Plant, Cell & Environment*, *42*(8), 2384–2398.

Terry, N. (1980). Limiting factors in photosynthesis: I. Use of iron stress to control photochemical capacity in vivo. *Plant Physiology*, *65*(1), 114–120.

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. (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.

Thaler, J. S., Humphrey, P. T., & Whiteman, N. K. (2012). Evolution of jasmonate and salicylate signal crosstalk. *Trends in Plant Science*, *17*(5), 260–270.

Theeuwen, T. P. J. M., Logie, L. L., Harbinson, J., & Aarts, M. G. M. (2022). Genetics as a key to improving crop photosynthesis. *Journal of Experimental Botany*, *73*(10), 3122–3137.

Thoen, M. P. M., Davila Olivas, N. H., Kloth, K. J., Coolen, S., Huang, P., Aarts, M. G. M., Bac-Molenaar, J. A., Bakker, J., Bouwmeester, H. J., & Broekgaarden, C. (2017). Genetic architecture of plant stress resistance: multi-trait genome-wide association mapping. *New Phytologist*, *213*(3), 1346–1362.

Tian, D., Araki, H., Stahl, E., Bergelson, J., & Kreitman, M. (2002). Signature of balancing selection in Arabidopsis. *Proceedings of the National Academy of Sciences*, *99*(17), 11525–11530.

Tian, Z., Rizzon, C., Du, J., Zhu, L., Bennetzen, J. L., Jackson, S. A., Gaut, B. S., & Ma, J. (2009). Do genetic recombination and gene density shape the pattern of DNA elimination in rice long terminal repeat retrotransposons? *Genome Research*, *19*(12), 2221–2230.

Tibbs Cortes, L., Zhang, Z., & Yu, J. (2021). Status and prospects of genome-wide association studies in plants. *The Plant Genome*, *14*(1), e20077.

Torres, M. A., & Dangl, J. L. (2005). Functions of the respiratory burst oxidase in biotic interactions, abiotic stress and development. *Current Opinion in Plant Biology*, *8*(4), 397–403.

Torres, M. A., Dangl, J. L., & Jones, J. D. G. (2002). Arabidopsis gp91phox homologues AtrbohD and AtrbohF are required for accumulation of reactive oxygen intermediates in the plant defense response. *Proceedings of the National Academy of Sciences*, *99*(1), 517–522.

Tsunoyama, Y., Ishizaki, Y., Morikawa, K., Kobori, M., Nakahira, Y., Takeba, G., Toyoshima, Y., & Shiina, T. (2004). Blue light-induced transcription of plastid-encoded psbD gene is mediated by a nuclear-encoded transcription initiation factor, AtSig5. *Proceedings of the National Academy of Sciences*, *101*(9), 3304–3309.

Turgut-Kara, N., Arikan, B., & Celik, H. (2020). Epigenetic memory and priming in plants. *Genetica*, *148*, 47–54.

Turner, D. P., Urbanski, S., Bremer, D., Wofsy, S. C., Meyers, T., Gower, S. T., & Gregory, M. (2003). A cross-biome comparison of daily light use efficiency for gross primary production. *Global Change Biology*, *9*(3), 383–395.

van Bezouw, R. F. H. M., Keurentjes, J. J. B., Harbinson, J., & Aarts, M. G. M. (2019). Converging phenomics and genomics to study natural variation in plant photosynthetic efficiency. *The Plant Journal*, *97*(1), 112–133.

Van de Peer, Y., Maere, S., & Meyer, A. (2009). The evolutionary significance of ancient genome duplications. *Nature Reviews Genetics*, *10*(10), 725–732.

Van de Peer, Y., Mizrachi, E., & Marchal, K. (2017). The evolutionary significance of polyploidy. *Nature Reviews Genetics*, *18*(7), 411–424.

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., & Thibault, J. (2013). From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics*, *43*(1), 10–11.

van Rooijen, R., Aarts, M. G. M., & Harbinson, J. (2015). Natural genetic variation for acclimation of photosynthetic light use efficiency to growth irradiance in Arabidopsis. *Plant Physiology*, *167*(4), 1412–1429.

Van Rooijen, R., Kruijer, W., Boesten, R., Van Eeuwijk, F. A., Harbinson, J., & Aarts, M. G. M. (2017). Natural variation of YELLOW SEEDLING1 affects photosynthetic acclimation of Arabidopsis thaliana. *Nature Communications*, *8*(1), 1421.

VanBuren, R., Wai, C. M., Ou, S., Pardo, J., Bryant, D., Jiang, N., Mockler, T. C., Edger, P., & Michael, T. P. (2018). Extreme haplotype variation in the desiccation-tolerant clubmoss Selaginella lepidophylla. *Nature Communications*, *9*(1), 13.

Vanneste, K., Baele, G., Maere, S., & Van de Peer, Y. (2014). Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous–Paleogene boundary. *Genome Research*, *24*(8), 1334–1347.

Veer, G. van der. (2006). *Geochemical soil survey of the Netherlands. Atlas of major and trace elements in topsoil and parent material; assessment of natural and anthropegenic enrichment factors* (Issue 347). Utrecht University.

Vernikos, G., Medini, D., Riley, D. R., & Tettelin, H. (2015). Ten years of pan-genome analyses. *Current Opinion in Microbiology*, *23*, 148–154.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., & Bright, J. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261–272.

Visscher, P. M., & Goddard, M. E. (2019). From RA Fisher’s 1918 paper to GWAS a century later. *Genetics*, *211*(4), 1125–1130.

Vlad, D., Kierzkowski, D., Rast, M. I., Vuolo, F., Dello Ioio, R., Galinha, C., Gan, X., Hajheidari, M., Hay, A., & Smith, R. S. (2014). Leaf shape evolution through duplication, regulatory diversification, and loss of a homeobox gene. *Science*, *343*(6172), 780–783.

von Caemmerer, S., & Evans, J. R. (2010). Enhancing C3 photosynthesis. *Plant Physiology*, *154*(2), 589–592.

Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., & Schatz, M. C. (2017). GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics*, *33*(14), 2202–2204.

Walter, J., & Kromdijk, J. (2022). Here comes the sun: How optimization of photosynthetic light reactions can boost crop yields. *Journal of Integrative Plant Biology*, *64*(2), 564–591.

Wang, J., Tao, F., Marowsky, N. C., & Fan, C. (2016). Evolutionary fates and dynamic functionalization of young duplicate genes in Arabidopsis genomes. *Plant Physiology*, *172*(1), 427–440.

Wang, X., Torres, M. J., Pierce, G., Lemke, C., Nelson, L. K., Yuksel, B., Bowers, J. E., Marler, B., Xiao, Y., & Lin, L. (2011). A physical map of Brassica oleracea shows complexity of chromosomal changes following recursive paleopolyploidizations. *BMC Genomics*, *12*(1), 1–15.

Wang, Y., Xiong, G., Hu, J., Jiang, L., Yu, H., Xu, J., Fang, Y., Zeng, L., Xu, E., & Xu, J. (2015). Copy number variation at the GL7 locus contributes to grain size diversity in rice. *Nature Genetics*, *47*(8), 944–948.

Waters, B. M., McInturf, S. A., & Stein, R. J. (2012). Rosette iron deficiency transcript and microRNA profiling reveals links between copper and iron homeostasis in Arabidopsis thaliana. *Journal of Experimental Botany*, *63*(16), 5903–5918.

Weaver, S., Dube, S., Mir, A., Qin, J., Sun, G., Ramakrishnan, R., Jones, R. C., & Livak, K. J. (2010). Taking qPCR to a higher level: Analysis of CNV reveals the power of high throughput qPCR to enhance quantitative resolution. *Methods*, *50*(4), 271–276.

Weigel, D. (2012). Natural variation in Arabidopsis: from molecular genetics to ecological genomics. *Plant Physiology*, *158*(1), 2–22.

Weigel, D., & Nordborg, M. (2015). Population genomics for understanding adaptation in wild plant species. *Annual Review of Genetics*, *49*, 315–338.

Werk, K. S., Ehleringer, J., Forseth, I. N., & Cook, C. S. (1983). Photosynthetic characteristics of Sonoran Desert winter annuals. *Oecologia*, *59*, 101–105.

Whale, A. J., King, M., Hull, R. M., Krueger, F., & Houseley, J. (2022). Stimulation of adaptive gene amplification by origin firing under replication fork constraint. *Nucleic Acids Research*, *50*(2), 915–936.

Wickham, H. (2016). *ggplot2: elegant graphics for data analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Wijfjes, R. Y., Smit, S., & De Ridder, D. (2019). Hecaton: reliably detecting copy number variation in plant genomes using short read sequencing data. *BMC Genomics*, *20*, 1–13.

Wilson, T. E., Arlt, M. F., Park, S. H., Rajendran, S., Paulsen, M., Ljungman, M., & Glover, T. W. (2015). Large transcription units unify copy number variants and common fragile sites arising under replication stress. *Genome Research*, *25*(2), 189–200.

Wu, A., Hammer, G. L., Doherty, A., von Caemmerer, S., & Farquhar, G. D. (2019). Quantifying impacts of enhancing photosynthesis on crop yield. *Nature Plants*, *5*(4), 380–388.

Wu, H.-J., Zhang, Z., Wang, J.-Y., Oh, D.-H., Dassanayake, M., Liu, B., Huang, Q., Sun, H.-X., Xia, R., & Wu, Y. (2012). Insights into salt tolerance from the genome of Thellungiella salsuginea. *Proceedings of the National Academy of Sciences*, *109*(30), 12219–12224.

Wu, J., & Baldwin, I. T. (2010). New insights into plant responses to the attack from insect herbivores. *Annual Review of Genetics*, *44*, 1–24.

Würschum, T., Boeven, P. H. G., Langer, S. M., Longin, C. F. H., & Leiser, W. L. (2015). Multiply to conquer: copy number variations at Ppd-B1 and Vrn-A1 facilitate global adaptation in wheat. *BMC Genetics*, *16*, 1–8.

Würschum, T., Longin, C. F. H., Hahn, V., Tucker, M. R., & Leiser, W. L. (2017). Copy number variations of CBF genes at the Fr-A2 locus are essential components of winter hardiness in wheat. *The Plant Journal*, *89*(4), 764–773.

Xiao, H., Jiang, N., Schaffner, E., Stockinger, E. J., & van der Knaap, E. (2008). A retrotransposon-mediated gene duplication underlies morphological variation of tomato fruit. *Science*, *319*(5869), 1527–1530.

Xie, M., Sun, J., Gong, D., & Kong, Y. (2019). The roles of Arabidopsis C1-2i subclass of C2H2-type zinc-finger transcription factors. *Genes*, *10*(9), 653.

Xu, C. A. I., Yinan, C. U. I., Zhang, L., Jian, W. U., Liang, J., Cheng, L., Xiaowu, W., & Cheng, F. (2018). Hotspots of independent and multiple rounds of LTR-retrotransposon bursts in Brassica species. *Horticultural Plant Journal*, *4*(4), 165–174.

Xu, W., Fu, W., Zhu, P., Li, Z., Wang, C., Wang, C., Zhang, Y., & Zhu, S. (2019). Comprehensive analysis of CRISPR/Cas9-mediated mutagenesis in arabidopsis thaliana by genome-wide sequencing. *International Journal of Molecular Sciences*, *20*(17), 4125.

Xu, Z., & Wang, H. (2007). LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research*, *35*(suppl_2), W265–W268.

Yadav, N. S., Titov, V., Ayemere, I., Byeon, B., Ilnytskyy, Y., & Kovalchuk, I. (2022). Multigenerational exposure to heat stress induces phenotypic resilience, and genetic and epigenetic variations in Arabidopsis thaliana offspring. *Frontiers in Plant Science*, *13*, 728167.

Yan, Z., Jing, M., Zhang, B., Shi, H., Jin, X., Yan, X., Gao, T., & Han, Y. (2023). The Arabidopsis LARP1s are Involved in Regulation of Seed Germination. *Journal of Plant Growth Regulation*, *42*(3), 1775–1788.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, *24*(8), 1586–1591.

Yao, Y., & Kovalchuk, I. (2011). Abiotic stress leads to somatic and heritable changes in homologous recombination frequency, point mutation frequency and microsatellite stability in Arabidopsis plants. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, *707*(1–2), 61–66.

Younginger, B. S., Sirová, D., Cruzan, M. B., & Ballhorn, D. J. (2017). Is biomass a reliable estimate of plant fitness? *Applications in Plant Sciences*, *5*(2), 1600094.

Youssef, A., Laizet, Y., Block, M. A., Maréchal, E., Alcaraz, J., Larson, T. R., Pontier, D., Gaffé, J., & Kuntz, M. (2010). Plant lipid-associated fibrillin proteins condition jasmonate production under photosynthetic stress. *The Plant Journal*, *61*(3), 436–445.

Yu, X., & Gabriel, A. (2003). Ku-dependent and Ku-independent end-joining pathways lead to chromosomal rearrangements during double-strand break repair in Saccharomyces cerevisiae. *Genetics*, *163*(3), 843–856.

Zaidem, M. L., Groen, S. C., & Purugganan, M. D. (2019). Evolutionary and ecological functional genomics, from lab to the wild. *The Plant Journal*, *97*(1), 40–55.

Zandalinas, S. I., Fichman, Y., & Mittler, R. (2020). Vascular bundles mediate systemic reactive oxygen signaling during light stress. *Plant Cell*, *32*(11), 3425–3435.

Zarrei, M., MacDonald, J. R., Merico, D., & Scherer, S. W. (2015). A copy number variation map of the human genome. *Nature Reviews Genetics*, *16*(3), 172–183.

Zhang, B., Jia, J., Yang, M., Yan, C., & Han, Y. (2012). Overexpression of a LAM domain containing RNA-binding protein LARP1c induces precocious leaf senescence in Arabidopsis. *Molecules and Cells*, *34*, 367–374.

Zhang, F., Gu, W., Hurles, M. E., & Lupski, J. R. (2009). Copy number variation in human health, disease, and evolution. *Annual Review of Genomics and Human Genetics*, *10*, 451–481.

Zhang, F., Khajavi, M., Connolly, A. M., Towne, C. F., Batish, S. D., & Lupski, J. R. (2009). The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. *Nature Genetics*, *41*(7), 849–853.

Zhang, H., Zhu, J., Gong, Z., & Zhu, J.-K. (2022). Abiotic stress responses in plants. *Nature Reviews Genetics*, *23*(2), 104–119.

Zhang, L., Cai, X., Wu, J., Liu, M., Grob, S., Cheng, F., Liang, J., Cai, C., Liu, Z., & Liu, B. (2018). Improved Brassica rapa reference genome by single-molecule sequencing and chromosome conformation capture technologies. *Horticulture Research*, *5*.

Zhang, L., & Jiménez-Gómez, J. M. (2020). Functional analysis of FRIGIDA using naturally occurring variation in Arabidopsis thaliana. *The Plant Journal*, *103*(1), 154–165.

Zhang, Y. I., & Turner, J. G. (2008). Wound-induced endogenous jasmonates stunt plant growth by inhibiting mitosis. *PloS One*, *3*(11), e3699.

Zhang, Z., Mao, L., Chen, H., Bu, F., Li, G., Sun, J., Li, S., Sun, H., Jiao, C., & Blakely, R. (2015). Genome-wide mapping of structural variations reveals a copy number variant that determines reproductive morphology in cucumber. *The Plant Cell*, *27*(6), 1595–1604.

Zhao, M., Du, J., Lin, F., Tong, C., Yu, J., Huang, S., Wang, X., Liu, S., & Ma, J. (2013). Shifts in the evolutionary rate and intensity of purifying selection between two Brassica genomes revealed by analyses of orthologous transposons and relics of a whole genome triplication. *The Plant Journal*, *76*(2), 211–222.

Zheng, L.-Y., Guo, X.-S., He, B., Sun, L.-J., Peng, Y., Dong, S.-S., Liu, T.-F., Jiang, S., Ramachandran, S., & Liu, C.-M. (2011). Genome-wide patterns of genetic variation in sweet and grain sorghum (Sorghum bicolor). *Genome Biology*, *12*(11), 1–15.

Zheng, X., Chen, L., Xia, H., Wei, H., Lou, Q., Li, M., Li, T., & Luo, L. (2017). Transgenerational epimutations induced by multi-generation drought imposition mediate rice plant's adaptation to drought condition. *Scientific Reports*, *7*(1), 39843.

Zheng, X., Gogarten, S. M., Lawrence, M., Stilp, A., Conomos, M. P., Weir, B. S., Laurie, C., & Levine, D. (2017). SeqArray—a storage-efficient high-performance data format for WGS variant calls. *Bioinformatics*, *33*(15), 2251–2257.

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.

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. (2017). Exploring structural variation and gene family architecture with De Novo assemblies of 15 Medicago genomes. *BMC Genomics*, *18*(1), 1–14.

Zhou, X., & Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*, *44*(7), 821–824.

Zhou, Z., Jiang, Y., Wang, Z., Gou, Z., Lyu, J., Li, W., Yu, Y., Shu, L., Zhao, Y., & Ma, Y. (2015). Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nature Biotechnology*, *33*(4), 408–414.

Zhu, J., Pearce, S., Burke, A., See, D. R., Skinner, D. Z., Dubcovsky, J., & Garland-Campbell, K. (2014). Copy number and haplotype variation at the VRN-A1 and central FR-A2 loci are associated with frost tolerance in hexaploid wheat. *Theoretical and Applied Genetics*, *127*, 1183–1197.

Zhu, X., Xie, S., Tang, K., Kalia, R. K., Liu, N., Ma, J., Bressan, R. A., & Zhu, J.-K. (2021). Non-CG DNA methylation-deficiency mutations enhance mutagenesis rates during salt adaptation in cultured Arabidopsis cells. *Stress Biology*, *1*, 1–12.

Zhu, X.-G., Long, S. P., & Ort, D. R. (2010). Improving photosynthetic efficiency for greater yield. *Annual Review of Plant Biology*, *61*, 235–261.

Zmienko, A., Marszalek-Zenczak, M., Wojciechowski, P., Samelak-Czajka, A., Luczak, M., Kozlowski, P., Karlowski, W. M., & Figlerowicz, M. (2020). AthCNV: a map of DNA copy number variations in the Arabidopsis genome. *The Plant Cell*, *32*(6), 1797–1819.

Żmieńko, A., Samelak, A., Kozłowski, P., & Figlerowicz, M. (2014). Copy number polymorphism in plant genomes. *Theoretical and Applied Genetics*, *127*, 1–18.

Zmienko, A., Samelak-Czajka, A., Kozlowski, P., Szymanska, M., & Figlerowicz, M. (2016). Arabidopsis thaliana population analysis reveals high plasticity of the genomic region spanning MSH2, AT3G18530 and AT3G18535 genes and provides evidence for NAHR-driven recurrent CNV events occurring in this location. *BMC Genomics*, *17*(1), 1–16

Summary

Plants are frequently exposed to various environmental conditions which can be stressful. However, environmental conditions that act as stress to one species, population or individual may not affect others, or may even be optimal to their fitness, as these may have acquired traits to better cope with such conditions. Adaptive evolution, the process through which beneficial genetic mutations are acquired, is a particularly important mechanism shaping plant adaptation. Genetic variation among organisms can lead to variation in heritable traits and consequently fitness. Understanding the genetic basis underlying plant adaptive traits, as well as the mechanisms that shape them, is important to develop more stress-tolerant crops.

Copy number variation (CNV), a type of genetic variation that involves the gain or loss of fragments of DNA in the genome, plays a significant role in driving phenotypic variation across various organisms (**Chapter 2**). CNVs are abundant in plants and exhibit dynamic changes in plant genomes over a relatively short evolutionary time. Initially, CNVs involving gene duplications are mostly redundant, primarily affecting the expression of the duplicated genes. However, over time, duplicated genes can acquire new functions or undergo alterations, losing their redundancy. CNVs are often associated with adaptation and evolution of plants in response to unfavorable environmental conditions. Understanding the mechanisms underlying CNV formation is crucial for comprehending plant evolution. However, detecting CNVs in plant genomes is challenging, and the development of new genome sequencing technologies and high-throughput validation methods is necessary to facilitate their detection and interpretation. Confirming their assumed high frequency will further emphasize their significance as a source of genetic variation that can be selectively utilized to enhance the creation of new superior crop cultivars.

In **Chapter 3**, the contribution of CNV to plant adaptation was evaluated by conducting an experiment in which plants were exposed to adverse environmental conditions. Using *Arabidopsis thaliana*, experimental evolution populations were created that were exposed to moderate and severe salinity and zinc stress for five generations. The populations exhibited noticeable physiological and phenotypic changes, resulting in reduced fitness due to the stress exposure. We identified spontaneous (epi)genetic mutations that occurred after five generations. Notably, plants subjected to severe zinc stress displayed a twofold higher mutation rate compared to the control group, while the salinity stress treatments maintained a stable mutation rate. Furthermore, we examined whether rapid adaptation took place in response to the salinity or zinc stress treatments. Interestingly, one of the replicate populations exposed to severe zinc stress demonstrated a significant improvement in performance under zinc stress conditions. These plants exhibited larger size and greatly enhanced fitness, as indicated by the increased number of seeds produced. Our findings suggest that this population adapted through a loss-of-function mutation in the *RBOHF* gene. *RBOHF* is a NADPH oxidase involved in generating reactive oxygen species and plays a crucial role in stress signaling pathways.

In **Chapter 4** a new *A. thaliana* natural variation panel was constructed, with accessions all coming from the Netherlands. Natural

populations of *A. thaliana* offer valuable insights into the adaptation of wild plant species. Previous research has predominantly focused on global populations or accessions collected from regions with diverse climates. However, the genetics underlying adaptation in regions with mild environmental clines remain poorly understood. In this study, we investigated a diversity panel consisting of 192 *A. thaliana* accessions collected from the Netherlands, a region with limited climatic variation. Despite the relative uniform climate, we identified compelling evidence of local adaptation within this population. Notably, semidwarf accession occur at a relatively high frequency near the coast and these displayed enhanced tolerance to high wind velocities. Additionally, we evaluated the performance of the population under iron deficiency conditions and found that allelic variation in the *FSD3* gene affects tolerance to low iron levels. Moreover, we explored patterns of local adaptation to environmental clines in temperature and precipitation, observing that allelic variation at *LARP1c* affects drought tolerance. Not only is the genetic variation observed in a diversity panel of *A. thaliana* collected in a region with mild environmental clines comparable to that in collections sampled over larger geographic ranges, it is also sufficiently rich to elucidate the genetic and environmental factors underlying natural plant adaptation.

Comparisons of closely related species which differ in their adaptive responses can be a powerful approach to study plant adaptative traits. In **Chapter 5**, an example of this is presented using *Hirschfeldia incana*, a member of the Brassicaceae family, which has exceptionally high rates of photosynthesis under high irradiance compared to the vast majority of other C_3-photosynthesizing plant species. While the core mechanisms of photosynthesis are highly conserved in C_3 plants, these mechanisms are very flexible, allowing considerable diversity in photosynthetic properties. Because *H. incana* is easy to grow, it is an excellent model for studying the genetic and physiological basis of this trait. By using comparative genomics we tested if interspecific CNV is an important component underlying the improved photosynthesis. The analysis showed for a set of photosynthesis genes that those genes that are retained at a higher copy number are more likely to be highly expressed. This may imply a functional contribution of those genes and may be a first clue into to the high photosynthesis of *H. incana*. Although this should be investigated further, the high-quality genome assembly will be an imperative resource to conduct further research on this trait.

S

In conclusion, this thesis has studied the genetic basis of plant adaptation using various approaches and has provided further insights into the role of CNV in plants. It has demonstrated that plant genetic adaptation to the environment can occur rapidly in highly adverse conditions and may involve higher mutation rates. Moreover, it showed that the use of natural variation can be a powerful tool to study the genes underlying responses to environmental stress. The establishment of the DartMap diversity panel and *Hirschfeldia incana* as a model system to study photosynthesis have already provided new insights into various plant adaptive traits, and will continue to do so.

Curriculum vitae

About the author

René Boesten was born on April 24, 1992 in Waalre, the Netherlands. He completed his BSc Biology at Wageningen University in 2013 with a thesis at the Laboratory of Genetics on signatures of selection in natural populations of the heavy metal hyperaccumulator plant *Noccaea caerulescens*. He then continued with the MSc Biology at Wageningen University, for which he did functional analysis of two photosynthesis-related genes in *Arabidopsis thaliana* during his MSc thesis, again at the Laboratory of Genetics. During his MSc, he did an internship at the Max Planck Institute for Plant Breeding Research where he used CRISPR/Cas9 to induce knockout mutations in several *SPL* transcription factors. After finishing his MSc studies (*cum laude*) in 2016, he started his PhD at the Laboratory of Genetics under the supervision of Prof. Dr. Mark Aarts and Prof. Dr. Bas Zwaan, of which the results are presented in this thesis. Since September 2021, he is working at the Laboratory of Genetics as a teacher in the education career path.

List of publications

Van Rooijen, R., Kruijer, W., **Boesten, R**., Van Eeuwijk, F. A., Harbinson, J., & Aarts, M. G. M. (2017). Natural variation of YELLOW SEEDLING1 affects photosynthetic acclimation of Arabidopsis thaliana. *Nature Communications*, *8*(1), 1421.

Stuster, R., Poelman E. H., & **Boesten, R.** (2017). Gifkikkers (*1st edition*). *Dendrobatidae Nederland.*

Garassino, F.*, Wijfjes, R. Y.*, **Boesten, R.***, Reyes Marquez, F., Becker, F. M., Clapero, V., Van den Hatert, I., Holmer, R., Schranz, M. E., Harbinson, J., De Ridder, D., Smit, S., & Aarts, M. G. M. (2022). *The Plant Journal*, 1125(5), 1298-1315.

Almira Casellas, M. J., Pérez-Martin, L., Busoms, S., **Boesten, R**., Llugany, M., Aarts, M. G. M., & Poschenrieder, C. (2023). A genome-wide association study identifies novel players in Na and Fe homeostasis in Arabidopsis thaliana under alkaline-salinity stress. *The Plant Journal, 113*(2), 225-245.

Stuster, R., Poelman E. H., & **Boesten, R.** (2023). Gifkikkers (*2nd edition*). *Dendrobatidae Nederland.*

Wijfjes, R. Y.*, **Boesten, R.***, Becker, F. M., Theeuwen, T. P. J. M., Snoek, B., Mastoraki, M., Verheijen, J. J., Güvencli, N., Denkers, L-A. M., Koornneef, M., Van Eeuwijk, F., Smit, S., De Ridder, D., & Aarts, M. G. M. (2023). Local adaptation of Arabidopsis thaliana in a small geographic region with mild environmental clines. *biorXiv.org* https://doi.org/10.1101/2023.09.18.558200

Boesten, R.*, Wijfjes, R. Y.*, Becker, F. M., Van der Woude, J., Van Workum, D-J. M., F., Smit, S., De Ridder, D., & Aarts, M. G. M. (2023). Rapid adaptation of Arabidopsi thaliana to excess zinc stress. *In preparation.*

** shared first authorship*

Acknowledgements

Early in the year 2013 I started my journey at the Laboratory of Genetics with my BSc Biology thesis under the supervision of Ya-Fen Lin and Mark. This brief, two-month thesis was quite an eye-opener to me. Mostly because up to that point I never took my education very seriously, as after all, a 6 is also sufficient. Yet, during the work discussions I quickly came to the realisation that in 2-3 years from then I would have to apply for a job somewhere. Someone was going to ask me: So René, what do you actually know? What can you do that is of any use to us? My honest answers to such questions were quite unsettling. Oopsie-woopsie.

Little did I know back then that this particular job application would turn out to be this PhD position. But as I had really enjoyed my short stay at the Laboratory of Genetics during my BSc thesis, I came back for my MSc thesis under the supervision of Roxanne van Rooijen and again Mark. Then, based off the recommendations by Mark and Maarten I pursued my internship at the Max Planck Institute for Plant Breeding, under the supervision of Youbong Hyun and George Coupland. After that, Mark and Roxanne offered me to work a few months as a student assistant on some experiments to finalize a paper, and Fons asked if I wanted to assist the Genetic Analysis Trends and Concepts course. Therefore, I really wanted to continue and was very happy to know that I obtained the position on which this thesis is about. This book does therefore not just marks the end of this thesis, but already more than 10 years (with a few breaks) of my stay at the Laboratory of Genetics. Thank you very much for making it such an enjoyable time. Much appreciated.

Mark, the text above already clearly demonstrates how important you have been throughout my entire development. Before applying to this project, I had attempted to obtain another PhD grant in the EPS graduate programme. It got rejected and I was wondering whether or not I should wait to try submitting it elsewhere. A bit later this project came available and I was initially quite sceptical towards the project. Significant adaptation within five generations? Yeah sure… I thought. But after you asked me a second time to have an actual look into the proposal, I became enthusiastic and applied. I’m very glad I did so. Thanks for all the years of support, opportunities and trust. Especially also for challenging me to do better at times, and sometimes to convince me to be a bit less critical. I have also always appreciated and enjoyed the barbecues, drinks and Sinterklaas evenings that you organised!

Bas, thank you for your advice and feedback, especially in the late stages of my thesis. Yet more importantly, I want to thank you for your constant effort in keeping a great atmosphere at our department at which I’ve always felt to be in my place. I appreciate your trust in me and that I have received the opportunity to do the work that I really like. Finally, I would also like to thank your fanatism during lab-outings. It really adds to the shininess of my Golden Medal of the Crazy 88 lab-outing.

Raúl, your great contribution to this thesis is more than obvious and clear by looking at the author lists of each chapter. I am very happy that we could work together on this project. Your calm and positive demeanour have always been helpful to me when times got more of a ‘challenge’. If it wasn’t for you, I doubt that I would have managed to send my presentation around in time even a single time to the rest of the team for one of the project meetings. Besides, I have (too) especially enjoyed our Wednesday lunches (independent from whether they involved either Wednesday or lunch). We had great discussions on how incredibly talented and gifted of a player Jorrit Hendrix is and how we were essentially like a monkey with cymbals.

I extent my gratitude to Sandra and Dick for your invaluable contributions to this thesis. Thank you for your insightful ideas and constructive suggestions during our NWO Groen meetings. I’m also particularly grateful for all of the feedback you have provided on my chapters and your support at some of the hardest moments during my PhD. I’m very happy that I’ve had the opportunity to closely collaborate with the three of you and I have learned a lot from this.

Frank, you had a large contribution to each of my experimental chapters with your ideas, solutions and loads of help with experiments. Most memorable to me are the prep days for the evolution experiment where we were preparing the ‘breads’ to go into the oven. I found a philosophical and inspirational quote from Cultureel Centrum ‘De Moefflon’ that captures the essence of our work perhaps best: “En jij was erbij toen water werd weggestemd, en jij was erbij toen vuur werd weggestemd. Dus jij zou kunnen be… Mag ik even uitspreken?! Jij zou ook kunnen begrijpen dat nadat water en vuur zijn weggestemd, en aarde natuurlijk sowieso uit den boze is, we op wind terecht zijn gekomen.” That ‘Thema Wind’ made it to this book is just great.

Maarten, thank you very much for all of your help in general and especially in regards to the chapter on the DartMap diversity panel. With over one third (68 out of 192) of the accessions used in the diversity panel that were collected by you, it is evident why we used to refer to the panel as the ‘Koornneefjes set’. Me (and also Raúl) have always very much appreciated your ideas, suggestions and feedback and this has made this chapter thereby a whole lot stronger.

Tom, Roel, Jitpanu and Ben I’m very happy that I’ve been able to do my PhD so closely together with the four of you. I received a lot of help and advice from each of you on different aspects of my project. But more importantly, you became friends and made this thing a whole lot more fun. Tom, I always appreciate our many discussions, no matter if they are about mapping genes, photosynthesis or the true meaning of wisdom and creativity. Let’s have many more to come. Roel, our sense of humour aligned well and I still miss those days where we were both sitting behind our computer, making ‘particular’ kind of noises (Ooh thank you, thank you). Also the Among Us nights where the others had already left are a great memory, but one that should not be shared. Jitpanu, or better just Panunu, you have been a true gem. The vrijmibo’s are simply not the same without our beloved Chipspanu. Also still thanks again for giving great advice on where to go in Thailand, even though I asked you only quite last-minute. I hope in the future I can visit you over there, ride the elephants into war and claim what is rightfully ours. Ben, you have been the most reliable vrijmibo member throughout all of those years. I'm happy that we were both so closely involved in teaching GATC and now share an office, so that our gossip doesn't necessarily have to wait until Friday.

Francesco, already during my PhD I met you for the first time during GATC. Only months later you started your MSc thesis in the only project where Raúl and me could jointly supervise a student on a, for us too, completely new project. The start of the Hirschfeldia incana project for which I'm happy that we could publish a paper together on this. You then managed to catch up on my 2+ years head start and we were both finishing up. During this time I really appreciated the talks we had while we were both struggling to get it all done in time. Your feedback on my propositions during these last stretches was certainly very helpful. And then you even graduated before me, congrats! By the way, when you read this, please give Pepa a nice treat for being such a good doggo.

I would also like to express my gratitude to all the colleagues from Mark's group who have helped me in different ways. Valeria, Yanli, Tania and Robert really helped me at the start to get going. Phoung, thanks for helping me on several occasions with further understanding GWAS. Laavanya, thanks for contributing your data to my Discussion. Sofia, for helping me with setting up the thesis ring. Henk, thanks for the suggestions on several experiments involving heavy metals. Alan, thanks for your ever-present smile and great Captain Cabbage acting skills. Louise, you have a great skill at making me feel very old. Thanks for all the funny conversations about non-relevant topics. Whether it is about your adventures at the fashion week (vroem vroem) or photosynthesis, I hope there will be many more to come. Finally, Jeroen and Ai, I'm glad that you both have now recently started on your PhD and I'm very much looking forward to the years to come.

The best way to put things into perspective at the end of the week has always been the vrijmibo. I'm very happy that you allowed me to have some weekly rants about something unimportant that happened, the definition of chairs or someone's propositions. Since many of the regulars who have made the vrijmibo a success have been mentioned (or will be) already, just a shout-out to Jenny, Mariska, Helena, Hylke, Maggie-Anne, Mirko, Nicky, Chris, Alex, Mariana, Spyros and Martin. Then, Alanna, thanks for the lovely Christmas cards over the years, and I find it admirable how you could take some of Jitpanu's roasts like such a champ. Murambiwa 'Buffalo Soldier' Nyati, it is a pity (to me) that you could not spend as much time in Wageningen, but it is always a pleasure when you are around. Bo, for several years Roel, Jitpanu, Tom and me got complaints from others about noise in the open work space. I'm happy that you and Louise have taken up the responsibility to fight the silence and continue our legacy.
I would also like to thank Wytske and Marjan for having the offices keep running smoothly and helping with countless questions that are so important

to get anything done. Wytske, thank you for all the support on all things surrounding my thesis, always good suggestions for fun lab activities and the many fun coffee breaks. Marjan, you make my life so much easier with how much work you take away from me. It's always nice and effective to work together, and I really appreciate your general kindness and positive attitude. Both of you have been/are key in making this a nice place to work.

The same holds true in regards to the lab, where the team of technicians continuously make sure that the lab is a safe and productive place. So many thanks to Bram, Corrie, Eric, Francisca, Frank, Gabriella, Gema, Jordy, José, Michiel and Patrick. Although I have always realised your great contribution, it has become even more clear to me ever since I joined the technicians meetings. There really is a lot more that you are constantly doing and arranging than I was previously aware of. Despite that the lab is, and will be, a constant flux of (new) people coming and going, I am very happy with your continuous patience towards everyone. It really provided me a nice and safe environment to learn and to conduct all my experiments.

Corrie, thank you for the many times you helped me in the seed lab, greenhouse and hydroponics room. Thanks for the many fun chats on the hallway and during the practicals of Plant Biotechnology and Plant Cell and Tissue Culture. I do however think you should join for Sinterklaas a bit more often. Francisca, besides having contributed directly to my thesis with several experiments, I also would like to thank you for helping so many of my students with their lab work. At times I wonder how you can still get around while also helping so many people at the same time. At times I think of you as an octopus with in each arm a pipette saving someone else's project.

For all of my experiments that involved growing plants, I'm grateful for the help I received from the different people working at Unifarm: Bert, Bertus, Gerrit, Taede, David, Rinie, Jannick, Sean, Martijn, Geurt, Rohan and Wilfred. In particular I want to thank Bert for your attentive eye on my experiments. Your care for my plants has really saved several experiments and improved many others.

Besides working on my PhD, I also spend considerable amount of time teaching in courses. Most of this time was spend in GATC. If it wasn't for this course, I would probably never have considered to do anything related to teaching. Fons you really changed my mind and perspective on this. Your constant effort and enthusiasm to make the course fun has been most important for that. As I think Tina Turner already sang in one of her songs, you are simply Debets! (badum tss)

GATC was always a nice change of setting and helped to take a step back from the everyday work and discuss my project with other colleagues in much more detail than is normally possible. These discussions, whether related to my project or about genetics or science in general, were a lot of fun and also very helpful. For this I would therefore especially like to thank some members of the (former) GATC team; Ben, Arjan, Eric Bastiaans, Erik 'Doppi-King' Wijnker, Joost K, Alex, Sabine, Suzette and Alan. In particular, Ben also thanks for the many nice discussions and gossip sessions

Besides teaching in courses, I have worked together and supervised many students. I enjoyed working together with you on projects and I am thankful for their invaluable contributions, whether in conducting experiments, analyzing data, or providing insightful discussions during meetings. Only mentioning your name here does not do justice to your contributions, but I hope you understand with the number of students which I have supervised during these years. Imme Bartels, Iris Boonstra, Feitse Bos, Pembe Canavar, Lissy-Anne Denkers, Francisco Garassino, Nuri Güvencli, Frank Hamer, Wouter Hijl, Hanna Hogenboom, Marco Jansen, Vindhya Jayachandran, Ioannis Kandylas, Nick Kersten, Fleur Kleijburg, Emile Klein-Gotink, Maria Mastoraki, Fariha Naz Apon, Tim Neefjes, Lanique Niels, Maxime Nigon, Anne-Fleur Peters, Mandy Ravensbergen, Angela Sagt, Su Sarlar, Jeroen van Buren, Anne van der Horst, Max van der Sandt, Jeroen van der Woude, Daniël Verbaan, Jelle Verheijen, Sietze Wals, Jun Wang, Ruyuan Wang, Chujie Yan and Yanda Zhou, thank you very much!

Ik wil ook graag mijn groep vrienden bedanken enzo. Wat was het toch gezellig hè! Ooh, echt gezellig. Jeetje. Nou, wel jammer dat ik soms net die ene dag niet kon. Jelle, bedankt voor alle Kara-vrijdagen en de verschillende propositie brainstormavonden. Jammer dat veel van die proposities niet echt controversieel waren, want verder hadden ze wel potentie. Reinier, ik ben blij dat ik met jou ten minste één persoon heb in de groep die zich wel af en toe gewoon volwassen gedraagt, net als ik. Judith/'t Grollemannetje/ Judipff, Ja. Wat was ~~Vietnam~~ Thailand toch mooi. Huehuehue. Jeroentje, we moeten echt weer een keer naar een reptielenbeurs, en dan voor elkaar een huisdier uitkiezen. Joren, samen met Sinan waren onze avonturen eindeloos. Capspert, van jou leerde ik dat oog voor detail, ook in namen, altijd telt. Beste Mosman, bedankt dat je je altijd in blijft zetten voor kikkertjes, erg belangrijk.

Als laatste wil ik nog graag mijn familie bedanken. Pap, mam, Berry, Liya, Emma en Willem. Bedankt voor alle goeie zorgen en dat jullie altijd voor me klaarstaan en me helpen waar mogelijk. Pap, bedankt dat je mij al op redelijk jonge leeftijd meenam naar Zuid-Afrika en later samen met Berry (of alleen met mij) die mooie reizen hebt gemaakt naar Costa Rica. Voor mij is er niets leukers dan lekker door de jungle struinen opzoek naar beestjes, en dat heeft een grote invloed gehad op mijn keuze om later biologie te gaan studeren (en daarin ben doorgegaan). Mam, ook bedankt dat ik al die jongere jaren, ondanks mijn allergieën voor allerlei harige dieren, toch zoveel andere leuke huisdieren mocht houden, ondanks dat vooral jij daar niet altijd zo'n fan van was. Verder bedank ik jullie voor alle hulp die ik overal continu bij heb gekregen al die jaren, van ophalen en wegbrengen, van alles wat jullie gedaan hebben in en rondom mijn huis (en eerder kamer) en de vele (soms ongewenste) 'aanmoedigingen' om dit boekje nu eindelijk eens af te maken.

Colofon

The research described in this thesis was financially supported by the Dutch Research Council NWO (ALWGR_1015.9).

Financial support from Wageningen University for printing this thesis is gratefully acknowledged.

Cover design:	Fenna Schaap | www.fennaschaap.nl
Thesis lay-out:	René Boesten
Printing:	Proefschriftmaken | www.proefschriftmaken.nl