universität
wien

DISSERTATION / DOCTORAL THESIS

Titel der Dissertation /Title of the Doctoral Thesis

„Genome evolution following major evolutionary transitions in the genus *Arabidopsis*“

verfasst von / submitted by

Robin Burns, BA

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Doctor of Philosophy (PhD)

Wien, 2021/ Vienna 2021

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 794 685 490
Dissertationsgebiet lt. Studienblatt / field of study as it appears on the student record sheet:	Molekular Biologie / Molecular Biology
Betreut von / Supervisor:	Prof. Dr. Magnus Nordborg

Acknowledgements

Magnus Nordborg — thank you for being a great supervisor and for the many meetings where I could bounce ideas around and really develop critical thinking skills. Polina Novikova — thank you for the guidance and the long nights in Ghent of hashing out ideas in between drinks, food and coding. I hope all my future collaborations are as fun and productive. Danièle Filiault — thank you for pushing me to think, present and discuss science in a productive manner and for the many chats about almost every topic. Joanna Guniš -- thank you for teaching me how to pipette, and essentially how to do everything else on the bench. Mayela Soto — thank you for being a great PhD companion to have. Every member of the Nordborg group (and many others in the GMI) that I've had the pleasure of knowing, all of you, thank you for the input, comments, jokes and all round fun environment to do research in.

My family and friends deserve a special thanks for the support over the years in various ways. Mum and Dad for the constant encouragement and motivation, making the trips from Wexford to Vienna to take me to der Oper, fancy dinners and trips to museums. Suzanne my twin and better half and David my wonderful brother for keeping me grounded and dropping everything at a moments notice if I needed it. My brilliant cousins Lisa, Eva, Alex, Naomi and many other family members. Sean, Jack, Deniz, and Lara for being brilliant friends to have, and thank you Sean especially for the proof reading. Jorge, Julien, Teddy, Catrin, Lily, Arun, Gökçe, Mike, Ranjith, Sophie and Klaus for the calls, dinners, drinks, walks, discussion, movies and baking attempts. Once again, everyone, thanks!

Introduction 5

Chapter 1 186

Chapter 2 44

Chapter 3 102

Discussion 136

Abstract

Plant genomes show incredible diversity in their size, order, composition, redundancy, chromosome number and ploidy. With the increasing availability of long-read whole genome sequences, we are beginning to understand just how dense the tip of the iceberg is. The work described in this thesis compares plant genomes both between and within different plant species in order to examine the effect of two major evolutionary transitions on genome evolution: polyploidy and changes in mating type.

The genus *Arabidopsis* holds a wealth of information into how genome evolution is shaped by major genomic transitions and challenges over short evolutionary time, making it an attractive model genus. These challenges include polyploidization or whole genome duplication (WGD), the evolution of self-compatibility, changes in genome size and chromosomal fusions — to name a few.

In the recent history of the *Arabidopsis* genus, many polyploid species have originated during the recent glacial periods. To study the process of polyploidization, we examined *A. suecica*, a young allopolyploid generated ~16 kya through the hybridization *A. thaliana* and *A. arenosa*. To understand genomic changes following polyploidization in *A. suecica* it is important to first understand the evolutionary history of the species. We found that the majority of genetic polymorphisms in *A. suecica* are shared with *A. thaliana* and *A. arenosa*, suggesting the origin of *A. suecica* consisted of multiple hybridization events between populations of the ancestral species, rather than a single unique event.

With an understanding of its origin in hand, we examined the evolution of allopolyploidy in *A. suecica*. In contrast to expectations, we found no evidence of any genome shock. The genome is not re-arranged, transposable elements (TEs) are not out of control and there is no subgenome dominance in gene expression. Instead, we found evidence for gradual adaptation towards polyploidy. Meiotic genes on the *A. thaliana* subgenome are up-regulated in their expression, possibly in order to avoid aneuploidy and genome rearrangements that are common in lab-generated synthetic *A. suecica*, and genes involved in photosynthesis are up-regulated on the *A. arenosa* subgenome, possibly in response to the new cytoplasmic environment, as plastids are maternally inherited from *A. thaliana*.

Over longer evolutionary time, the genus has experienced dramatic changes to karyotype and genome size. A classic example is the shrunken ~125Mb genome of *A. thaliana* with 5 chromosomes compared to the ancestral genome size of ~200Mb with 8 chromosomes, which are the characteristics of the other *Arabidopsis* species, such as *A. lyrata* and the outgroup *Capsella*. By examining the genomes from multiple individuals and species, we found the shrinking of the *A. thaliana* genome to have involved multiple regions that appear unstable, the instability of these regions has resulted in multiple allelic variants of different sizes that are segregating both within and between the species.

Zusammenfassung

Pflanzengenome zeigen eine unglaubliche Vielfalt in ihrer Größe, Ordnung, Zusammensetzung, Redundanz, Chromosomenzahl und Ploidie. Mit der zunehmenden Verfügbarkeit von Long-Read-Ganzgenomsequenzen beginnen wir zu verstehen, wie dicht die Spitze des Eisbergs ist. Die in dieser Arbeit beschriebene Arbeit vergleicht Pflanzengenome sowohl zwischen als auch innerhalb verschiedener Pflanzenarten, um die Auswirkung von zwei großen evolutionären Übergängen auf die Genomevolution zu untersuchen: Polyploidie und Veränderungen im Paarungstyp.

Die Gattung *Arabidopsis* birgt eine Fülle von Informationen darüber, wie die Genomevolution durch große genomische Übergänge und Herausforderungen über kurze evolutionäre Zeiträume geprägt wird, was sie zu einer attraktiven Modellgattung macht. Zu diesen Herausforderungen gehören Polyploidisierung oder Ganzgenomduplikation (WGD), die Evolution der Selbstkompatibilität, Veränderungen der Genomgröße und chromosomale Fusionen - um nur einige zu nennen.

In der jüngeren Geschichte der Gattung *Arabidopsis* sind viele polyploide Arten während der letzten Eiszeiten entstanden. Um den Prozess der Polyploidisierung zu studieren, haben wir *A. suecica* untersucht, ein junges Allopolyploid, das ~16 kya durch die Hybridisierung von *A. thaliana* und *A. arenosa* entstanden ist. Um die genomischen Veränderungen nach der Polyploidisierung in *A. suecica* zu verstehen, ist es wichtig, zunächst die Evolutionsgeschichte der Art zu verstehen. Wir fanden heraus, dass die Mehrheit der genetischen Polymorphismen von *A. suecica* mit *A. thaliana* und *A. arenosa* geteilt werden, was darauf hindeutet, dass der Ursprung von *A. suecica* aus mehreren Hybridisierungsereignissen zwischen Populationen der angestammten Arten bestand und nicht aus einem einzigen Ereignis.

Mit dem Wissen um den Ursprung von *A. suecica* untersuchten wir die Evolution der Allopolyploidie in *A. suecica*. Im Gegensatz zu den Erwartungen fanden wir keine Hinweise auf einen Genomschock. Das Genom ist nicht neu arrangiert, transponierbare Elemente (TEs) sind nicht außer Kontrolle geraten und es gibt keine Subgenom-Dominanz in der Genexpression. Stattdessen fanden wir Hinweise auf eine allmähliche Anpassung in Richtung Polyploidie. Meiotische Gene auf dem *A. thaliana*-Subgenom sind in ihrer Expression hochreguliert, möglicherweise um Aneuploidie und Genom-Rearrangements zu vermeiden, die in der im Labor erzeugten synthetischen *A. suecica* üblich sind, und Gene, die an der Photosynthese beteiligt sind, sind auf dem *A. arenosa*-Subgenom hochreguliert, möglicherweise als Reaktion auf die neue zytoplasmatische Umgebung, da Plastiden mütterlicherseits von *A. thaliana* vererbt werden.

Über einen längeren evolutionären Zeitraum hat die Gattung dramatische Veränderungen des Karyotyps und der Genomgröße erfahren. Ein klassisches Beispiel ist das geschrumpfte ~125Mb Genom von *A. thaliana* mit 5 Chromosomen im Vergleich zur angestammten Genomgröße von ~200Mb mit 8 Chromosomen, die die anderen *Arabidopsis*-Arten, wie *A. lyrata* und die Außengruppe *Capsella*, aufweisen. Durch die Untersuchung der Genome von mehreren Individuen und Arten fanden wir heraus, dass die Schrumpfung des *A. thaliana*-Genoms mehrere Regionen betroffen hat, die instabil erscheinen. Die Instabilität dieser Regionen hat zu mehreren allelischen Varianten unterschiedlicher Größe geführt, die sowohl innerhalb als auch zwischen den Arten segregieren.

Introduction

In a letter to Joseph Hooker in 1879, Charles Darwin described the global radiation and general evolutionary success of angiosperms (flowering plants) as an "abominable mystery"[1]. Much like angiosperms, the field of genomics has also evolved and rapidly radiated over a relatively short period of time. One of the promises from the era of genomics is to use whole-genome sequencing to understand the evolution of genomes, and how they compare and contrast to other genomes from related species. As the field advances and costs decrease we are beginning to tackle more complex questions from larger and more diverse genomes.

In brief, plant genomes evolve through repeated cycles of changes in ploidy (i.e. the number of copies of chromosomes in a cell), the expansion and contraction of repeat content (such as transposable elements, centromeres and ribosomal DNA) and changes to the arrangement of gene order and gene copy number (i.e. gene duplication and gene loss). Other factors that influence genome evolution largely do so by changing how selection can act on the variation that exists or arises spontaneously *de novo* in a genome. These factors include: environmental changes, population bottlenecks, domestication, polyploidization, hybridization, and changes to a plants mating system (e.g. self-compatible vs. self-incompatible plants).

As we continue to investigate plant genome evolution, the great diversity and remarkable array of evolutionary patterns is becoming ever clearer, and long held ideas are being overturned. Unlocking these genetic secrets to describe model plant systems is becoming increasingly important in our need to improve crop breeding in agriculture and predict potential impacts of climate change on plant diversity. One factor, however, remains constant -- the "abominable mystery" continues to intrigue geneticists, ecologists, botanists, gardeners and plant enthusiasts of today.

A history of polyploidy throughout angiosperm evolution

Early on, and before the field of genomics experienced its rapid radiation, classic cytology studies on plant speciation reported that polyploidy or whole genome duplication (WGD), is a common mutation in angiosperms[2]. These observations were also independently supported by the study of fossil guard cells (epidermal cells that control the opening and closing of stomata on the leaf) from woody angiosperms. As cell size correlates positively with genome size, the genome size of the extinct woody angiosperms could be estimated, and ploidy level inferred[3]. As such, it was concluded that most angiosperms share polyploidy ancestry in their evolutionary history and, furthermore, that many exist as extant polyploids.

Also evident from the cytology studies was that the extant polyploids can be classified into one of two main types. The first is autopolyploidy, in which the duplicated genome belongs to the same individual. This can happen as a consequence of cell division errors, such as chromosome nondisjunction[4], a failure of the sister chromatids to separate properly in meiosis. Autopolyploidy can also involve the hybridization between closely related populations of the same species. The second is allopolyploidy which involves genome duplication via a hybridization step between two distinct species. The hybridization step that leads to polyploidy (auto- or allo-) likely involves the fusion of unreduced gametes, and this is regarded as a main mechanism of polyploid formation[5]. Unreduced gametes are gametes (e.g. a sperm or egg cell) that have not experienced the normal meiotic cell cycle process of reductive division. Unreduced gametes contain the full genetic information of an individual rather than the normal half. The two types of polyploids are illustrated in Figure 1

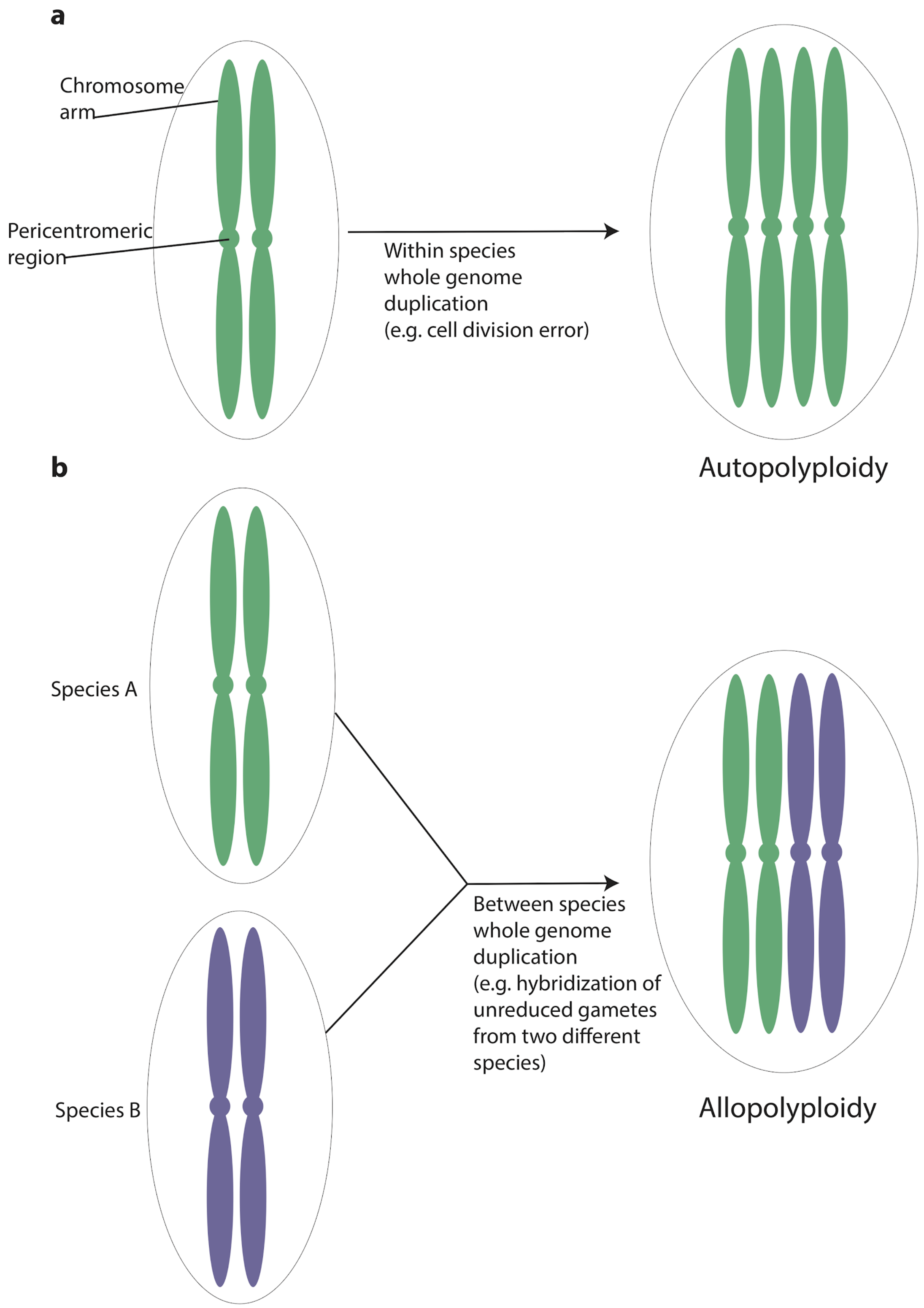

a
Chromosome arm
Pericentromeric region
Within species whole genome duplication (e.g. cell division error)
Autopolyploidy
b
Species A
Species B
Between species whole genome duplication (e.g. hybridization of unreduced gametes from two different species)
Allopolyploidy

Figure 1. The two main types of polyploidy are **a** autopolyploidy and **b** allopolyploidy.

Genomic studies over the last ~20 years have confirmed these observations and extended them to include that all angiosperms share a polyploid common ancestor, with many branches experiencing additional and subsequent rounds of polyploidy throughout their evolution[6]. We now know that the genomes of angiosperms have (and still do) cycle through polyploid and diploid states. From the available angiosperm species that have been examined to date ~33% are believed to be current polyploids[7]. It is therefore likely that polyploidy has played an important role in the radiation of angiosperms and likely continues to today. This is also evident in the non random genome tracts that have been left over from polyploidy[8]. It is also true in our own genomes[9]. For example, the four Hox gene clusters, that play an important role in early development, are relics that are left over from the two successive rounds of whole genome duplication at the base of the vertebrate tree[10]. Genomics, therefore, has redrawn the angiosperm tree to include polyploidy both in the past and present sense. Both the extent and prevalence of polyploidy in angiosperms, compared to other organisms such as animals, is striking due to the many challenges that are faced by any new polyploid species. The reason why polyploidy is rarer in animals than in plants is itself not understood, but likely involves the constraint of sex chromosome dosage in animals, and the fact that many plants are hermaphroditic which can allow for sex to occur despite a change in cytotype.

Population genetic challenges faced by new polyploid species

Polyploidization can be thought to represent a bona fide and instantaneous speciation event, as it results in an individual that is reproductively isolated from its diploid relatives, mainly due to the barrier of having a difference in chromosome number[11]. Interploidy crosses will likely result in deleterious aneuploid individuals. One of the main challenges faced by any new polyploid species is the loss of genetic variation through a severe genetic bottleneck. A genetic bottleneck, as the name suggests, is a term used to describe loss of genetic variation by demographic processes. This can occur as a result of a drastic reduction in a population's size (e.g. after a prolonged drought period), or a new speciation event (e.g. polyploidy). Genetic bottlenecks reduce the efficiency of selection acting on a population and increase the probability of an allele to reach fixation (i.e. present in all individuals) or be lost (removal from the population) by chance alone. Changes in allele frequencies by chance from changes in demographics is referred to as genetic drift. For example, alleles that are slightly deleterious are more likely to reach high frequency when the size of the population is small. In addition, beneficial alleles that are present in just a few individuals at the start of a bottleneck or arise de novo during the bottleneck are also more likely to be lost in a small population. In both cases loss or fixation depends on the strength of selection acting on an allele, and its initial starting frequency in the population. In a large population with genetically diverse individuals and many mates to choose from, the probability for loss or fixation by drift alone is less likely, and many alleles are available for selection to act on.

Counteracting a loss of variation through genetic bottlenecks, new polyploid species may backcross to their diploid progenitors. This would be difficult for allopolyploids, as crossing to either diploid progenitor will result in chromosome pairing for only one of the subgenomes, likely resulting in deleterious aneuploid combinations. However, gene flow between diploid and tetraploid populations in the autopolyploid *A. arenosa* have been reported. While no evidence for triploid individuals have yet been found in nature[12] an interesting hypothesis is that diploid *A. arenosa* individuals produce unreduced gametes at a high enough rate and that these gametes can (being the correct ploidy level to do so) fuse with the gametes from tetraploid *A. arenosa* individuals[13]. Despite the limitation for gene flow from progenitors in allopolyploids, allos do have fixed heterozygosity because of the hybridization event involved. However, heterozygosity can be a double edged sword, and depending on how compatible the alleles on the different subgenomes are, the relationship can result in either hybrid vigor[14] or hybrid necrosis[15].

A final important way in which polyploids may be alleviated from loss of genetic diversity is if many individuals from populations of the progenitors contribute to the origin (i.e. a unique vs. multiple origin). Multiple origins could provide a large enough allele pool for selection to act on, making extinction of the new polyploid species less likely. Therefore, when studying polyploid genomes it is important to first understand the demographic history of the species and its relation to its progenitors, in order to understand how selection may act on the given (sub-)genome.

Cell biology challenges faced by new polyploid species

Doubling the genome content has a dramatic effect on basic cell biology processes. These include cell division, recombination, protein folding, interactions with plastids, epigenetic changes, regulation of gene expression and transposable element (TEs) mobilisation. This is reviewed in greater detail in[16–18].

A distinguishing feature between autos and allos is how the chromosomes behave during cell division. Allopolyploids exhibit disomic inheritance. The two subgenomes, despite co-existing in the same cell, genetically behave as two diploid genomes and exchange little genetic material through recombination. Subgenomes preferentially pair during cell division and are bivalent (i.e. homologous chromosomes associate in pairs). Autopolyploids have tetrasomic inheritance. The duplicated chromosomes are interchangeable and randomly associate during cell division. A big challenge for autopolyploids is to avoid the formation of multivalents (instead of bivalents) during cell division as this will lead to failures in the faithful segregation of chromosomes. Evidence for selection for bivalent pairing has been reported in tetraploid *A. arenosa* which forms stable bivalents more frequently than its diploid *A. arenosa* relatives, which were artificially induced to be tetraploid as a comparison. While there is no evidence for preferential pairing in *A. arenosa*, a reduction in the number of crossovers from two to one crossover per bivalent likely contributed to a stable cell division phenotype[19].

In addition to the increased DNA content, allopolyploids also face the challenge of ensuring that two diverged genomes function together in a cell. As such, the harmonious picture painted in the previous paragraph of two diploid genomes working side by side in allos is in fact much more complex. The consequences to genome merging in allopolyploids will be here grouped under the umbrella term of genome shock. Genome shock and examples will be discussed further in Chapter 2 of this thesis and therefore will be briefly summarised here. The concept of genome shock is mentioned by Barbara McClintock in her Nobel lecture[20], and lays out a hypothesis of a unique set of challenges that genome merging presents a species with. This challenge can be thought of as any stress of genomic conflict between the genomes of two given species that is exposed by their hybridization. These conflicts are likely as the two subgenomes have been evolving independently for sometimes millions of years, before coming together again through hybridization. One

consequence likely is a dramatic rearrangement between the subgenomes of the allopolyploid[21], possibly because of stress on the cell division machinery and its ability to function correctly to ensure bivalent chromosome pairing for each subgenome.
Another challenge associated with genome shock is the mobility of TEs. TEs are selfish genetic elements which make copies of themselves through different mechanisms in the genome. Increased TE mobility is likely the consequence of conflicts in the TE silencing machinery[22] between the subgenomes and/or TE sequences coming into contact with a subgenome which may be naïve to it. The final challenge to do with genome shock I will discuss in this chapter relates to gene expression changes. As mentioned, plant genomes over evolutionary time have gone through cycles of poylploid and diploid states. Reverting back to a diploid state involves a genome fractionation process that is termed re-diploidization, however the process of re-diploidization and factors that influence it remain unclear. In autopolyploids, gene loss or fractionation may show no real bias as each gene is identical or very similar to its copy. In allopolyploids this likely shows a different pattern. The term subgenome dominance refers to a state in a hybrid or allopolyploid where one of the subgenomes is dominant in its expression over the other (i.e. genes are transcribed primarily from this genome). As the process of re-diploidization begins, the more dominant or expressed subgenome is likely to remain intact over the others, resulting in a pattern of biased gene loss. Predictions of which subgenome is likely to be dominant in expression, mainly are based on which subgenome has fewer TEs and thus less heterochromatin[23]. Whether the level of heterochromatin in a genome influences gene expression in euchromatin, however, is unclear.

In summary, new polyploid species face a lot of challenges both from a population genetics point of view and from a view of basic cell biology. However, given their abundance, current polyploids in angiosperms offer the unique research opportunity to study polyploidization "in action" and will help us answer some of these big questions.

Transitions from outcrossing to selfing

"Nature abhors self fertilization", is another well known statement from Darwin[24], and indeed Darwin connected reductions in vigor and fertility to inbreeding in plants[25]. Nevertheless, many flowering plants have undergone the major mating type transition from self-incompatibility (a.k.a. outcrossing) to self-compatibility (a.k.a. selfing).

The S locus system found in many plant species contains two tightly linked genes, the self recognition gene classified as S-LOCUS CYSTEINE-RICH PROTEIN (SCR) and S-LOCUS RECEPTOR KINASE (SRK). The SCR encodes a cysteine rich protein expressed on the pollen coat and is a ligand for the SRK, a receptor kinase expressed on the surface of the stigma. The locus is very polygenic and through interaction of SCR and SRK, outcrossing plants reject their own pollen or pollen from a plant that carries the same allele, while selfing plants can accept their own pollen, having lost this ability of recognition. Transitions to selfing have occurred multiple times and independently in many angiosperms[26]. In Brassicaceae, this includes recent allopolyploids such as *A. kamchatica, C. bursa pastoris* and *A. suecica* and diploids such as *C. rubella, C. orientalis*, and the workhorse of plant genetics *A. thaliana*. The evolutionary success of selfing plants has been attributed to having the benefit of reproduction assurance outweighing the cost of inbreeding depression. For a new cytotype, such as in allopolyploids, or when a plant is geographically isolated (e.g. while expanding into new habitats), being able to guarantee reproduction is advantageous. Selfing plants have also been shown to have a greater chance to become naturalized[27].

Like polyploidy, the evolutionary transition to selfing likely contributed to the evolutionary success of some flowering plant species, by opening up new possibilities for radiations into new habitats or just being able to avoid going extinct when no potential mates or required pollinators are available. Like polyploidy too, the consequences of selfing on genome evolution are not well understood. There is an ascertainment bias too for the assembled plants genomes to be selfing plants. This can be due to technical ease of assembling homozygous genomes and use as model species. However, large differences in the repeat content between selfers and their outcrossing congeners exist[28,29], and often selfing plants appear to have a shrunken genome[26,29,30]. This is also true for animal genomes, for example in *Caenorhabditis* nematodes[31].

As more genomes increasingly become available, and the types of genomes we can examine expands, answering this question is of interest to understand the evolution of plant genomes and their success. One way in which this is possible is the use of long-read sequencing as it has the ability to read through repetitive regions of the genome that are difficult to assemble.

Long read sequencing

New advances in sequencing technology are spurring new questions to be asked (and vice versa). Long read sequencing offers the advantage of reading through large repetitive regions of the genome that can assist greatly in genome assembly, and the longer reads allow for the easier distinguishment between subgenomes in allopolyploids. Two technologies are currently spearheading the advancement: Pacific BioSciences (PacBio) and Oxford Nanopore Technologies (ONT). PacBio platforms (e.g. Sequel II) use single molecule real time sequencing (SMRT) that involves examining the fluorescence patterns unique to each nucleotide that are added on by a polymerase tethered to the bottom of a sequencing well. While Nanopore sequencing (e.g. MinION) involves the read out of electric currents when nucleotides (single stranded) pass through a nanopore. Read length is more limited in PacBio as it relies on the polymerase activity, but are on average 20-30Kb, while Nanopore can have read lengths up to megabases in size[32]. The accuracy of the long reads for both technologies has increased since they started, with estimates ~1% for PacBio and ~5% error rates for Nanopore[32]. One difficulty for genome assembly that is not resolved by long reads, however, is the problem of heterozygosity.

Heterozygosity in genome assembly leads to bubbles structures to be present in the genome graph, where multiple paths are likely. Overcoming this problem, genome assemblers typically remove one of the paths in the bubble by pairwise alignment, however, this approach only works for sequences with low heterozygosity, and is likely inefficient for more heterozygous genomes which have clusters of structural variants, repeats and SNPs. One path forward to assembling heterozygous genomes, is to partition allelic variation before attempting a genome assembly. This can be done by using the trio binning approach, wherein haplotype information from parents in a pedigree are used to partition long reads generated from an F1 individual and can lead to a complete diploid genome assembly for the F1 individual[33]. Another path forward is to manually generate homozygous genomes from heterozygous individuals. This can be achieved by inbreeding the heterozygous individual over many generations, but inbreeding depression by the combination of recessive lethal alleles is a big bottleneck and depending on the generation time of the organism of interest, the approach can also be quite labour intensive and slow. A final approach to overcome heterozygosity in genome assembly is to generate double-haploids. Double haploid generation involves the artificial doubling of haploid cells (such as a gamete cell like pollen cells) to produce a homozygous diploid. This technique has been used in plant-breeding to decrease labor time of crosses and has started to be used to aid genome assembly such as in pear[34] and also in pufferfish[35].

The model genus *Arabidopsis*

While angiosperms contain an array of beautiful systems to work on, the genus *Arabidopsis* stands out with its simplicity, flexibility and diversity in answering questions in plant genetics and genome evolution. The genus contains closely related species that differ in a multitude of ways including their ecology, chromosome number, genome size, mating type and ploidy. The species are also widely distributed geographically across the northern hemisphere. The most well known is the annual plant *A. thaliana* which differs from its sister species by having a smaller genome size (~125Mb compared to the ancestral ~200Mb), a reduction from 8 to 5 chromosomes and is selfing plant. *A. thaliana* last shared a common ancestor with its relatives ~6 Mya[29]. *A. thaliana* also has a large number of resources available, including a broad sampling of its natural variation (i.e. the 1001 genomes project), a well assembled and annotated reference genome and multiple phenotypes measured that are publically available for performing GWAS. The rest of the genus likely diverged ~1Mya and includes the outcrossing perennials *A. lyrata* (that also has an available reference genome assembly), *A. halleri* and *A. arenosa* and the more geographically restricted *A. cebennensis* and *A. pedemontana*. The genus also contains recent polyploids that likely originated during the recent glacial periods. The connection between the glacial periods and polyploidization events may be tied to cold environments increasing the occurrence of unreduced gametes, such that these gametes were abundant and could fuse to produce polyploid plants. The polyploids of *Arabidopsis* include autopolyploid populations of *A. lyrata* and *A. arenosa* and the allopolyploids *A. kamachatic*a (a hybrid between *A. lyrata* and *A. halleri*) and *A. suecica*, (a hybrid between *A. thaliana* and *A. arenosa*)[36].

A. suecica is a model allopolyploid whose evolution is discussed in this thesis in the first two chapters. The genetic divergence (~12%) of its progenitors *A. arenosa* and the model plant *A. thaliana*, together make *A. suecica* an attractive system to study genome shock as well as providing ease in distinguishing subgenomes in genome assembly. Allopolyploids are typically more common in nature than autopolyploids. However, it is important to note that this result may be because of a bias in the ability to distinguish allos more readily from their progenitors than autos, because the hybridization event involved may produce more distinguishing phenotypes in allos. Allopolyploids are also abundant in cultivated crops, and these include cotton, strawberries, oat, wheat and peanuts. Maize and soy (which are now genetically diploid plants) also have experienced very recent allopolyploidization events. Studying the evolution of current polyploids in nature can provide valuable insight into the reasons behind their evolutionary success and also help separate the impacts of polyploidy versus domestication on the genomes and the phenotypes of agriculturally important crop species.

In the third chapter of this thesis, a comparative genomics approach is used to compare genome size evolution following mating type transitions in *Arabidopsis*. As reference genomes exist for *A. thaliana* and *A. lyrata* they provide a suitable model system to understand genome size differences. In North America, outcrossing populations of *A. lyrata* have low genetic diversity and have experienced population expansions following the last Ice Age northward towards the Great Lakes of North America[37]. The populations of *A. lyrata* in North America can genetically be divided into East and West populations, and at the margin of each population (but not the interior) selfing populations of *A. lyrata* have evolved[38]. The existence of selfing *A. lyrata* populations allows for an easier genome assembly and is a suitable proxy for outcrossing *Arabidopsis*, given the short evolutionary time since the last Ice Age.

Aims

The aims of this thesis are to present a comprehensive analysis of genome evolution following two major transitions: polyploidy and mating type. By using the *Arabidopsis* genus we are presented with the possibility to study both transitions and furthermore examine genome evolution on two different timescales, the first being since the last glacial maximum and the recent origin of new polyploid species and the second being over long evolutionary time since the divergence of *A. thaliana* from the other *Arabidopsis* species, that was likely followed by its transition to selfing[39].

The first two chapters relate to genome evolution following polyploidy and examining the allopolyploid *A. suecica.* With whole genome sequencing we first trace its evolutionary history and follow this with an analysis of the genomic consequences of polyploidy on genome structure and function. To do this, we assemble a chromosome level long read genome for *A. suecica* and use polymorphism and transcriptome data for multiple individuals from *A. suecica*, its progenitors, and synthetic *A. suecica* generated de novo in a lab environment.

The third chapter relates to genomic consequences following transitions in the reproductive mode or mating type of plants. To investigate this, we examine the evolution of genome shrinkage associated with a transition to self fertilization in *Arabidopsis*. By taking a comparative species approach and by using long read genome assemblies from multiple individuals, we examine positive and negative length variants between the genomes of *A. thaliana* and *A. lyrata* and ask the question to what extent does genome shrinkage involve *de novo* mutations vs selection acting on standing variation?

By using the *Arabidopsis* genus we gain insight into genome evolution and speciation on two different evolutionary time points. Over a shorter evolutionary time ~16Kya, we learn of the evolution of allopolyploidy in *A. suecica*. Examining genome evolution over longer evolutionary time ~6Mya, we learn how the genomes of the different *Arabidopsis* species are evolving since their most recent common ancestor. Linking these two time points together, we further our understanding of factors that lead to the divergence between species genomes and, at the same time, learn how these diverged genomes can hybridize and function in the same cell, allowing for the evolution of new successful plant species.

References

1. Darwin, C. *More Letters of Charles Darwin: A Record of His Work in a Series of Hitherto Unpublished Letters*. (1903).
2. Stebbins, G. L., Jr. Types of polyploids; their classification and significance. *Adv. Genet.* **1**, 403–429 (1947).
3. Masterson, J. Stomatal size in fossil plants: evidence for polyploidy in majority of angiosperms. *Science* **264**, 421–424 (1994).
4. Shi, Q. & King, R. W. Chromosome nondisjunction yields tetraploid rather than aneuploid cells in human cell lines. *Nature* **437**, 1038–1042 (2005).
5. Bretagnolle, F. & Thompson, J. D. Gametes with the somatic chromosome number: mechanisms of their formation and role in the evolution of autopolyploid plants. *New Phytol.* **129**, 1–22 (1995).
6. Jiao, Y. *et al.* Ancestral polyploidy in seed plants and angiosperms. *Nature* **473**, 97–100 (2011).
7. Rice, A. *et al.* The global biogeography of polyploid plants. *Nat Ecol Evol* **3**, 265–273 (2019).
8. Blanc, G. & Wolfe, K. H. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell* **16**, 1667–1678 (2004).
9. Lander, E. S. Initial impact of the sequencing of the human genome. *Nature* **470**, 187–197 (2011).
10. Hokamp, K., McLysaght, A. & Wolfe, K. H. The 2R hypothesis and the human genome sequence. *J. Struct. Funct. Genomics* **3**, 95–110 (2003).
11. Köhler, C., Mittelsten Scheid, O. & Erilova, A. The impact of the triploid block on the origin and evolution of polyploid plants. *Trends Genet.* **26**, 142–148 (2010).
12. Kolář, F. *et al.* Ecological segregation does not drive the intricate parapatric distribution of diploid and tetraploid cytotypes of theArabidopsis arenosagroup (Brassicaceae): Cytogeography ofArabidopsis arenosa. *Biol. J. Linn. Soc. Lond.* **119**, 673–688 (2016).
13. Monnahan, P. *et al.* Pervasive population genomic consequences of genome duplication in Arabidopsis arenosa. *Nat Ecol Evol* **3**, 457–468 (2019).
14. Chen, Z. J. Molecular mechanisms of polyploidy and hybrid vigor. *Trends Plant Sci.* **15**, 57–71 (2010).
15. Bomblies, K. & Weigel, D. Hybrid necrosis: autoimmunity as a potential gene-flow barrier in plant species. *Nat. Rev. Genet.* **8**, 382–393 (2007).

16. Parisod, C. *et al.* Impact of transposable elements on the organization and function of allopolyploid genomes. *New Phytol.* **186**, 37–45 (2010).

17. Doyle, J. J. & Coate, J. E. Polyploidy, the nucleotype, and novelty: The impact of genome doubling on the biology of the cell. *Int. J. Plant Sci.* **180**, 1–52 (2019).

18. Bomblies, K. When everything changes at once: finding a new normal after genome duplication. *Proc. Biol. Sci.* **287**, 20202154 (2020).

19. Yant, L. *et al.* Meiotic adaptation to genome duplication in Arabidopsis arenosa. *Curr. Biol.* **23**, 2151–2156 (2013).

20. McClintock, B. The significance of responses of the genome to challenge. *Science* vol. 226 792–801 (1984).

21. Chester, M. *et al.* Extensive chromosomal variation in a recently formed natural allopolyploid species, Tragopogon miscellus (Asteraceae). *Proc. Natl. Acad. Sci. U. S. A.* **109**, 1176–1181 (2012).

22. Liu, B. & Wendel, J. F. Epigenetic phenomena and the evolution of plant allopolyploids. *Mol. Phylogenet. Evol.* **29**, 365–379 (2003).

23. Steige, K. A. & Slotte, T. Genomic legacies of the progenitors and the evolutionary consequences of allopolyploidy. *Curr. Opin. Plant Biol.* **30**, 88–93 (2016).

24. Darwin, C. *On the Various Contrivances by which British and Foreign Orchids are Fertilised by Insects*. (John Murray, 1877).

25. Darwin, C. The effects of cross and self fertilization in the vegetable kingdom. John Murray, London. *The effects of cross and self fertilization in the vegetable kingdom. John Murray, London.* (1876).

26. Wright, S. I., Ness, R. W., Foxe, J. P. & Barrett, S. C. H. Genomic Consequences of Outcrossing and Selfing in Plants. *Int. J. Plant Sci.* **169**, 105–118 (2008).

27. Razanajatovo, M. *et al.* Plants capable of selfing are more likely to become naturalized. *Nature Communications* vol. 7 (2016).

28. Slotte, T. *et al.* The Capsella rubella genome and the genomic consequences of rapid mating system evolution. *Nat. Genet.* **45**, 831–835 (2013).

29. Hu, T. T. *et al.* The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nat. Genet.* **43**, 476–481 (2011).

30. Roessler, K. *et al.* The genome-wide dynamics of purging during selfing in maize. *Nat Plants* **5**, 980–990 (2019).

31. Yin, D. *et al.* Rapid genome shrinkage in a self-fertile nematode reveals sperm competition proteins. *Science* **359**, 55–61 (2018).

32. Amarasinghe, S. L. *et al.* Opportunities and challenges in long-read sequencing data analysis. *Genome Biol.* **21**, 30 (2020).

33. Koren, S. *et al.* De novo assembly of haplotype-resolved genomes with trio binning. *Nat. Biotechnol.* (2018) doi:10.1038/nbt.4277.

34. Linsmith, G. *et al.* Pseudo-chromosome–length genome assembly of a double haploid 'Bartlett' pear (Pyrus communis L.). *Gigascience* **8**, (2019).

35. Zhang, H. *et al.* Dramatic improvement in genome assembly achieved using doubled-haploid genomes. *Scientific Reports* vol. 4 (2015).

36. Novikova, P. Y. *et al.* Sequencing of the genus Arabidopsis identifies a complex history of nonbifurcating speciation and abundant trans-specific polymorphism. *Nat. Genet.* **48**, 1077–1082 (2016).

37. Lucek, K., Hohmann, N. & Willi, Y. Postglacial ecotype formation under outcrossing and self-fertilization in Arabidopsis lyrata. *Mol. Ecol.* **28**, 1043–1055 (2019).

38. Griffin, P. C. & Willi, Y. Evolutionary shifts to self-fertilisation restricted to geographic range margins in North American Arabidopsis lyrata. *Ecol. Lett.* **17**, 484–490 (2014).

39. Tang, C. *et al.* The evolution of selfing in Arabidopsis thaliana. *Science* **317**, 1070–1072 (2007).

Chapter1

This chapter was published in Molecular, Biology and Evolution, doi: 10.1093/molbev/msw299.

Genome sequencing reveals the origin of the allotetraploid *Arabidopsis suecica*

Polina Yu. Novikova[1,2], Takashi Tsuchimatsu[1*], Samson Simon[3], Viktoria Nizhynska[1], Viktor Voronin[1], Robin Burns[1], Olga M. Fedorenko[4], Svante Holm[5], Torbjörn Säll[6], Elisa Prat[7], William Marande[7], Vincent Castric[3], Magnus Nordborg[1†]

[1]Gregor Mendel Institute, Austrian Academy of Sciences, Vienna Biocenter (VBC), A-1030 Vienna, Austria.
[2]Vienna Graduate School of Population Genetics, Institut für Populationsgenetik, Vetmeduni, Vienna, Veterinärplatz 1, 1210, Wien, Austria
[3]CNRS / Université de Lille - Sciences et Technologies, Lille, France
[4]Institute of Biology, Karelian Research Center of the Russian Academy of Sciences, 185910 Petrozavodsk, Russia.
[5]Faculty of Science, Technology and Media, Department of Natural Sciences, Mid Sweden University, SE-851 70 Sundsvall, Sweden.
[6]Department of Biology, Lund University, 223 62 Lund, Sweden.
[7]Centre National de Ressources Génomiques Végétales, INRA-CNRGV, 24 Chemin de Borde Rouge, CS 52627, 31326 Castanet-Tolosan, France.
*Current address: Department of Biology, Chiba University, Yayoi-cho 1-33, Inage, Chiba 263-8522, Japan.
†Corresponding author. E-mail: magnus.nordborg@gmi.oeaw.ac.at

Keywords: Polyploidy, Arabidopsis suecica, Arabidopsis thaliana, Arabidopsis arenosa, shared polymorphism

Polyploidy is an example of instantaneous speciation when it involves the formation of a new cytotype that is incompatible with the parental species. Because new polyploid individuals are likely to be rare, establishment of a new species is unlikely unless polyploids are able to reproduce through self-fertilization (selfing), or asexually. Conversely, selfing (or asexuality) makes it possible for polyploid species to originate from a single individual — a *bona fide* speciation event. The extent to which this happens is not known. Here, we consider the origin of *Arabidopsis suecica*, a selfing allopolyploid between *A. thaliana* and *A. arenosa*, which has hitherto been considered to be an example of a unique origin. Based on whole-genome re-sequencing of 15 natural *A. suecica* accessions, we identify ubiquitous shared polymorphism with

the parental species, and hence conclusively reject a unique origin in favor of multiple founding individuals. We further estimate that the species originated after the last glacial maximum in Eastern Europe or central Eurasia (rather than Sweden, as the name might suggest). Finally, annotation of the self-incompatibility loci in *A. suecica* revealed that both loci carry non-functional alleles. The locus inherited from the selfing *A. thaliana* is fixed for an ancestral non-functional allele, whereas the locus inherited from the outcrossing *A. arenosa* is fixed for a novel loss-of-function allele. Furthermore, the allele inherited from *A. thaliana* is predicted to transcriptionally silence the allele inherited from *A. arenosa*, suggesting that loss of self-incompatibility may have been instantaneous.

Introduction

Polyploidy requires a series of unlikely events: the formation of unreduced gametes, hybridization, and the establishment of a new polyploid population (Ramsey 1998, Soltis, Marchant et al. 2015). Nevertheless, whole-genome duplication events have occurred throughout evolutionary history, and have been frequent in plants (Vision, Brown et al. 2000, Jiao, Wickett et al. 2011, Vanneste, Baele et al. 2014).

The genus *Arabidopsis* includes two relatively young allotetraploid species: *A. kamchatica* and *A. suecica* (N. 1957, Shimizu, Fujii et al. 2005, Shimizu-Inatsugi, Lihova et al. 2009). The former is a hybrid between *A. lyrata* and *A. halleri* and is limited to East Asia and North America (Shimizu, Fujii et al. 2005); the latter is a hybrid between *A. thaliana* and *A. arenosa* and is limited to the Fennoscandinavian region (O'Kane, Schaal et al. 1996). Previous studies have suggested that *A. suecica* originated from a single hybridization event between 12 Kya and 300 Kya (Jakobsson, Hagenblad et al. 2006) with *A. thaliana* as the maternal parent (Price 1994, Hurka 1995, Comai, Tyagi et al. 2000, Sall, Jakobsson et al. 2003). The latter conclusion is based partly on sequences from maternally inherited chloroplast genomes, partly on the fact that "synthetic" allotetraploids can be generated by fertilizing autotetraploid *A. thaliana* (which occur rarely in nature, but can readily be generated in the laboratory) with pollen from naturally autotetraploid *A. arenosa* (which are common), whereas the reciprocal cross cannot be made (Comai, Tyagi et al. 2000). Thus, the most likely scenario for the formation of *A. suecica* is that in which a normal (diploid) pollen from tetraploid *A. arenosa* fertilizes an unreduced gamete of diploid *A. thaliana* (Jakobsson, Hagenblad et al. 2006). In support of this scenario, the *A. arenosa* complement of *A. suecica* is more closely related to tetraploid rather than diploid *A. arenosa* (Novikova, Hohmann et al. 2016). The alternative scenario of mating between diploid parents followed by whole-genome duplication seems less likely.

The behavior of allotetraploid genomes during meiosis depends largely on the divergence between the parental species, because it allows to prevent homeologous pairing, but it also can be controlled by specific molecular mechanisms (Griffiths, Sharp et al. 2006). Cytological studies revealed a diploid-like, homologous chromosomal pairing in *A. suecica* (Comai, Tyagi et al. 2003). Interestingly, synthetic lines appear to be much less stable (Comai, Tyagi et al. 2000, Madlung, Tyagi et al. 2005, Henry, Dilkes et al. 2014). It has been suggested that such meiotic regularity has a genetic basis and is under selection in polyploids (Henry, Dilkes et al. 2014).

A. suecica is currently widely used as a model for studying allotetraploidy in terms of the evolutionary retention of homoeologs (Chang, Dilkes et al. 2010), the epigenetic regulation of nucleolar dominance (Chen, Comai et al. 1998, Pikaard 1999, Pontes, Lawrence et al. 2007, Costa-Nunes, Pontes et al. 2010, Pontvianne, Blevins et al. 2012), overall gene expression (Wang, Tian et al. 2006, Ha, Lu et al. 2009, Ng, Miller et al. 2014, Ng, Shi et al. 2014, Tian, Li et al. 2014, Miller, Song et al. 2015), and heterosis (Solhaug, Ihinger et al. 2016). One of the main advantages of *A. suecica* as a model (in addition to the fact that one of the parents is the model plant *A. thaliana*), is the possibility to “re-run evolution” by creating synthetic hybrids (Chen, Comai et al. 1998, Comai, Tyagi et al. 2000). However, to fully capitalize on this, it is important to understand the history and origin of the natural species better: hence this paper. Using whole genome sequencing data of multiple natural *A. suecica* accessions that cover most of its geographic distribution, we aim to describe the population history of this allotetraploid species: the location and timing of its origin and also the evolution of its ability to self-fertilize which ultimately led to the establishment of *A. suecica* as a new species.

Results and Discussion

We sequenced (using Illumina 100bp paired-end reads) 15 natural accessions of *A. suecica* sampled at different locations throughout the species distribution (supplementary table S1, supplementary fig. S1). We mapped *A. suecica* reads to the *A. thaliana* and *A. lyrata* reference genomes simultaneously, and obtained variant calls from the *A. thaliana* and *A. arenosa* components of *A. suecica*, respectively (see Materials and Methods). Our approach was greatly facilitated by the fact that *A. suecica* accessions are natural inbred lines as a result of selfing: by only retaining homozygous calls, we avoid many spurious polymorphisms that would have arisen from the misalignment of reads to the wrong parental genome. We mapped 77 percent of the raw reads on average (supplementary table S1), identifying 167,283 polymorphic sites in 15 *A. suecica* accessions on the *A. thaliana* portion of the reference and 416,898 sites on the *A. lyrata* portion. Throughout the study, results generated for *A. suecica* are compared with data from the parental species *A. thaliana* (Consortium 2016) and *A. arenosa* (Novikova, Hohmann et al. 2016).

Previous results have suggested that *A. suecica* had a unique origin, undergoing an extreme bottleneck that completely wiped out ancestral polymorphism, at least in the *A. thaliana* portion of the genome (note that a unique origin involving a selfing parent probably automatically eliminate most polymorphism; see Jakobsson, Hagenblad et al. 2006). We were thus very surprised to find that 89% of identified polymorphisms for the *A. thaliana* portion of *A. suecica* are shared with contemporary *A. thaliana*. A similar result was obtained for the *A. arenosa* portion of the genome: 91% of polymorphic sites are shared with the parental species (fig. 1, supplementary fig. S2). This amount of shared or, rather, retained ancestral variation clearly contradicts the previously suggested unique origin of *A. suecica* (Jakobsson, Hagenblad et al. 2006), especially since *A. thaliana* was already selfing when it contributed to *A. suecica* (see below, supplementary fig. S3), and thus is unlikely to have contributed more than one allele at each locus. Indeed, the genome-wide allele sharing between *A. suecica* and *A. thaliana* is incompatible with even a single generation of selfing in a putative single founder on the *A. thaliana* side (and, as we shall see below, many regions of the genome harbor more than two ancestral haplotypes and must therefore have more than a single ancestor).

Nevertheless, there are clear traces of a major bottleneck, presumably associated with the origin of the new polyploid species from a relatively small number of founders. Not only is the overall level of polymorphism strongly reduced (to roughly 30 and 12 percent of that of *A. thaliana* and *A. arenosa*, respectively), but also non-synonymous and putatively deleterious alleles are present at higher frequencies than in the parental species (supplementary fig. S4

A-B) — as expected as a consequence of drift during a bottleneck. We note, however, that purifying selection appears to have been operating after the establishment of the species: among polymorphisms private to *A. suecica* (*i.e.*, polymorphisms that must have arisen in the species), non-synonymous ones are again biased towards rare alleles (supplementary fig. S4 C-D).

There are also large chromosomal regions almost devoid of variation (fig. 1) in the *A. suecica* genome. While some of these may reflect selective sweeps in the new species, a simpler explanation is that they are a consequence of the foundation bottleneck. Indeed, the ubiquity and size of these regions will make it very difficult to find any genuine selective sweeps. The largest region, on the second chromosome of the *A. thaliana* portion of the genome, covers most of the long arm (~8 Mb). We can use this to estimate how old *A. suecica* is. Under the assumption that the small amount of polymorphism that does exist in this region has been generated solely by new (i.e., non-ancestral) mutations (4.5% of polymorphism in this genomic region is shared with *A. arenosa*, which should be compared with the genome-wide average of 91%, see above), we estimate that the bottleneck occurred ~ 16 Kya (95% CI [14.1 Kya, 18.4 Kya]: other bottlenecked regions give similar results; see *Materials and Methods* and below). Consistent with this, estimates of the effective population size over time (using MSMC; (Schiffels and Durbin 2014)) based on the full data point to a sharp decline following the last glacial maximum roughly 22 Kya (supplementary fig. S5). The decline is particularly noticeable in the *A. arenosa* portion of the genome, which is expected given that the ancestral species was an obligate outcrosser, implying that this portion underwent a transition to selfing as well.

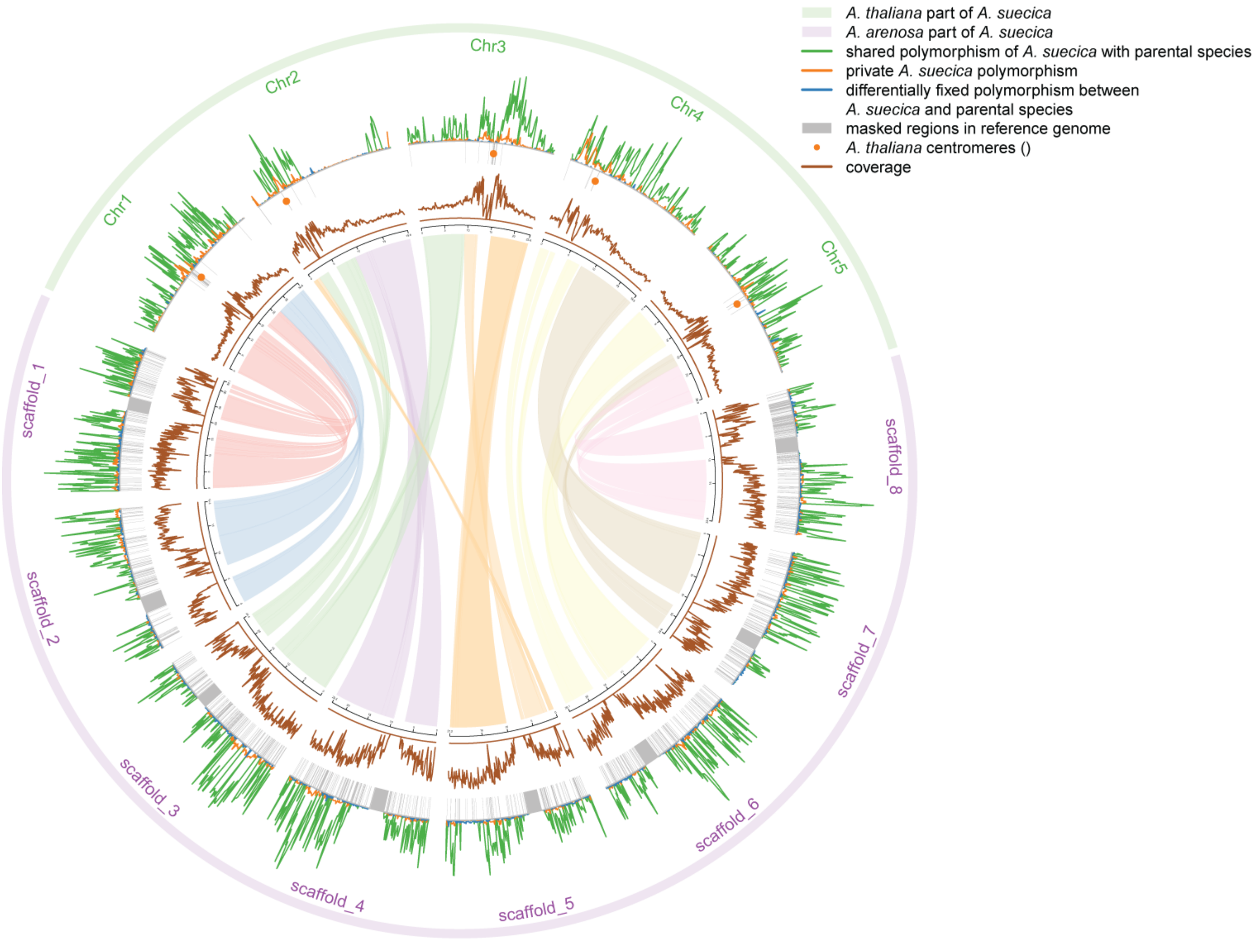


Fig. 1. Polymorphism density (outer graph) and sequencing coverage (inner, brown graph) along the chromosomes of *A. suecica* (shown on the outer rim, with the five *A. thaliana* chromosomes indicated in green and the eight *A. lyrata* reference genome scaffolds indicated in purple). The polymorphism density (number of SNPs per aligned site) along the genome is shown separately for shared (green), private (orange), and differentially fixed polymorphism (blue). A large non-polymorphic region is located between 7.8 Mbp and 16.2 Mbp on chromosome 2 of the *A. thaliana* portion of *A. suecica*. Links between the *A. thaliana* and *A. lyrata* reference genomes (center) are adapted from (Hu, Pattyn et al. 2011).

The next question is where *A. suecica* originated. We sought the ancestral *A. thaliana* population in the worldwide collection of sequenced *A. thaliana* genomes (Consortium 2016). Based on pairwise sequence divergence, the most closely related *A. thaliana* accessions appear to be found around the Ural Mountains, and in northern and central Eurasia (fig. 2). Clustering accessions using ADMIXTURE (Alexander, Novembre et al. 2009), similarly groups *A. thaliana* accessions from northern and central Eurasia with *A. suecica,* suggesting a shared past (supplementary fig. S6). Thus *A. suecica* is not most closely related to the *A. thaliana* with which it currently coexists, and the ancestral population must have been elsewhere. Indeed, because the Fennoscandinavian region was covered by ice until ~6 Kya (fig. 2; http://worldclim.org/paleo-climate; (Hijmans 2005), it is obvious that both species must be recent immigrants. Given that Swedish *A. thaliana* (unlike Swedish *A. suecica*) do not appear to be particularly closely related to Russian *A. thaliana* (Consortium 2016), a plausible scenario is that *A. thaliana* mainly reached Scandinavia from the south, via present-day Denmark (Consortium 2016), while *A. suecica* took the northern route, via present-day Finland.

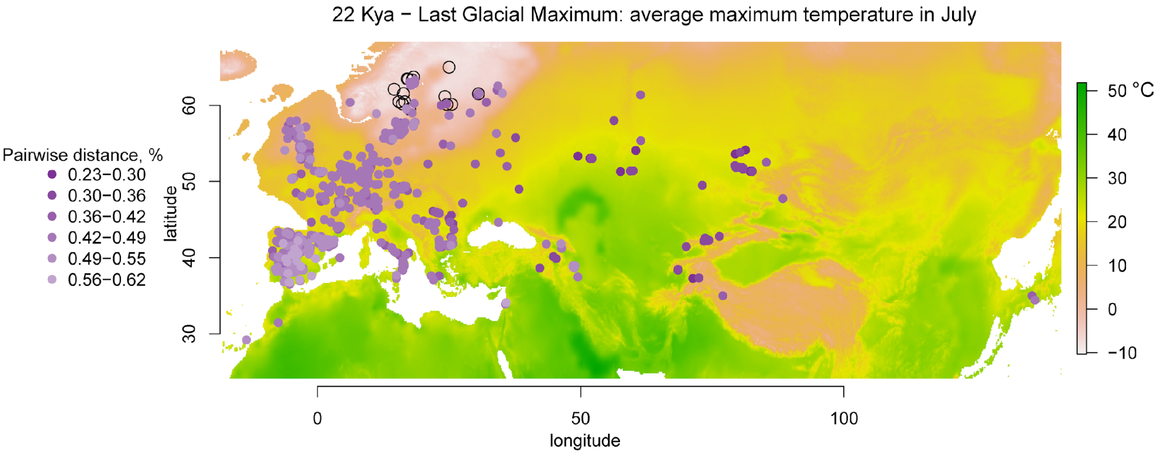


Fig. 2. The *A. thaliana* accessions most closely related to *A. suecica* are found in northern and central Eurasia (indicated by the dark violet circles). The background color of the map indicates the average maximum temperature in July during the last glacial maximum (Hijmans 2005). The present distribution of *A. suecica* (Fennoscandinavia; *A. suecica* sampling locations are indicated with open black circles) was covered by ice during this time, experiencing temperatures below zero degrees in July.

Interestingly, although *A. suecica* is clearly most closely related to Russian *A. thaliana* (fig. 2), the largest shared haplotype was found in *A. thaliana* from northern Sweden. An almost ~4 Mb segment of the ~8 Mb bottlenecked region on chromosome 2 (fig. 1) appears to be shared (i.e. identical-by-descent) with an *A. thaliana* accession from northern Sweden (fig. 3). However, this accession is not particularly closely related to *A. suecica* in any other sense, and the bottlenecked region does not show a pattern of relatedness different from the genome-wide pattern (fig. 3A-C). Since recent admixture is extremely unlikely (*A. thaliana* and *A. suecica* do not produce fertile offspring), one explanation is that this is a case of ancestral haplotype sharing, with the extreme length in northern Sweden being another facet of the generally much more extensive haplotype sharing and linkage disequilibrium in *A. thaliana* in this part of the world (Long, Rabanal et al. 2013). Simply put, there has effectively been less recombination in northern Swedish *A. thaliana* since the last glaciation, and this has led to greater haplotype sharing with ancient *A. thaliana*, and hence with *A. suecica*. Consistent with this interpretation, there is extensive haplotype sharing in this chromosomal region throughout Europe (fig. 3A).

Alternatively, the extensive haplotype sharing in northern accessions is a consequence of some *A. thaliana* having migrated to Sweden together with *A. suecica* via a northern route, and since admixed with *A. thaliana* coming from the south. Interestingly, the *A. thaliana* accession closest to *A. suecica* (by average pairwise divergence) is from the Altai mountain range in Central Asia (fig. 3B), a region which served as a refugium for many species (including humans (Reich, Green et al. 2010, Prufer, Racimo et al. 2014)) during the glacial interchanges (Tarasov 2000, Pavelkova Ricankova, Robovsky et al. 2014). One can thus imagine a scenario wherein a local *A. thaliana* population, which contributed to the formation of *A. suecica,* occupied new territories following the retraction of the ice sheets alongside its hybrid progeny. Where this would have taken place is far from clear. While the closest relatives on the *A. thaliana* side are currently found in Central Asia, there is no evidence for the other parent of *A. suecica* – *A. arenosa* – in this region. Thus, the most likely scenario may be that *A. suecica* originated somewhere in Eastern Europe and migrated to Fennoscandinavia following the retracting ice, while its parental *A. thaliana* population additionally spread into Central Asia.

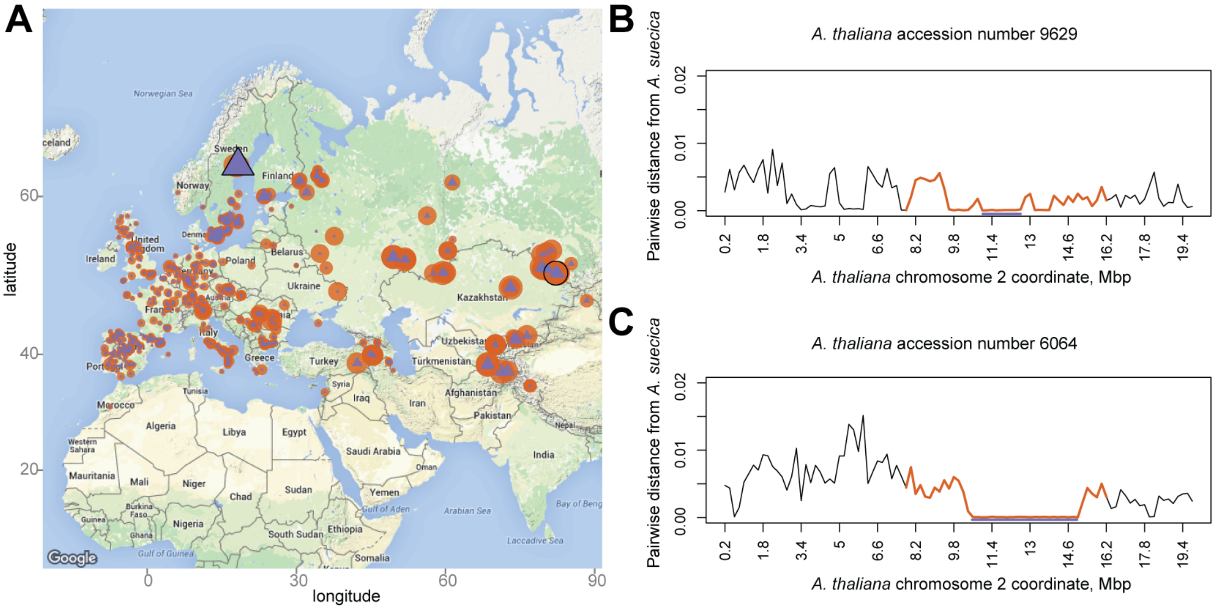


Fig. 3. The relationship between *A. suecica* and *A. thaliana* accessions in the bottlenecked region corresponding to *A. thaliana* chromosome 2 (fig. 1). (**A**) The size of the orange circles is inversely proportional to the average pairwise distance between *A. thaliana* and *A. suecica* w.r.t. the bottlenecked region of interest; this distribution is similar to the genome-wide pattern shown in fig. 2. The closest accession (9629) is highlighted with a black circle. The size of the violet triangles is proportional to the length of a haplotype that is shared with *A. suecica*. The accession with the longest haplotype (6064) is highlighted with a black triangle. (**B**,**C**) The average pairwise distance between 9629 and 6064, respectively, and *A. suecica*. The violet line shows the position of the longest haplotype that is shared

with *A. suecica*; the orange lines show the region inherited from one founder in *A. suecica* and used to calculate the pairwise distance on fig. 3A.

Next, we considered the number of *A. thaliana* individuals that contributed to the founding of *A. suecica*. As we have seen, the number of founding haplotypes is one for several regions of the genome, most noticeably on chromosome 2 (fig. 1 and fig. 3). As it happens, another example (as in fig. 4A, see below) is a region located at the top of chromosome 4, which harbors the four loci previously used to conclude that *A. suecica* likely had a unique origin (Jakobsson, Hagenblad et al. 2006). This conclusion was thus correct for this region, but is clearly incorrect for most of the genome. To gain further insight into the number of founders, we searched the genome for ancestral haplotype blocks shared between *A. suecica* and *A. thaliana* (using PLINK; see *Materials and Methods*). We identified 1273 haplotype blocks for which all *A. suecica* accessions fell into a single cluster of almost identical haplotypes, which we interpret as sharing a single founder haplotype (fig. 4A). Similarly, we found 1267 blocks for which the accessions can be clearly divided into 2 haplotype clusters; 106 for which they can be divided into three haplotype clusters, and two for which there were four clusters (fig. 4B-D). As a consistency check, we estimated the divergence time for haplotypes belonging to the same founder haplotype and obtained a very similar estimate to that reported above for the chromosome 2 region (95% CI [15.1 Kya, 16.6 Kya], see *Materials and Methods*).

The observed numbers of founder haplotypes do not directly correspond to the number of *A. thaliana* founders, as lineages may have been lost through drift (Nordborg 1998). Of course the number of haplotypes provides a lower bound, and we can therefore conclude that at least four founding individuals contributed (under the assumption that the founders were inbred, which is likely). To see if it is possible to be more precise, we simulated gene genealogies under linear growth models with varying numbers of founding individuals at the estimated time of origin (~16 Kya). These simulations showed that the observed distribution of founder lineages is compatible with a wide range of parameters (supplementary fig. S7), and it is therefore unlikely that we will be able to refine our estimate of the number of founders further.

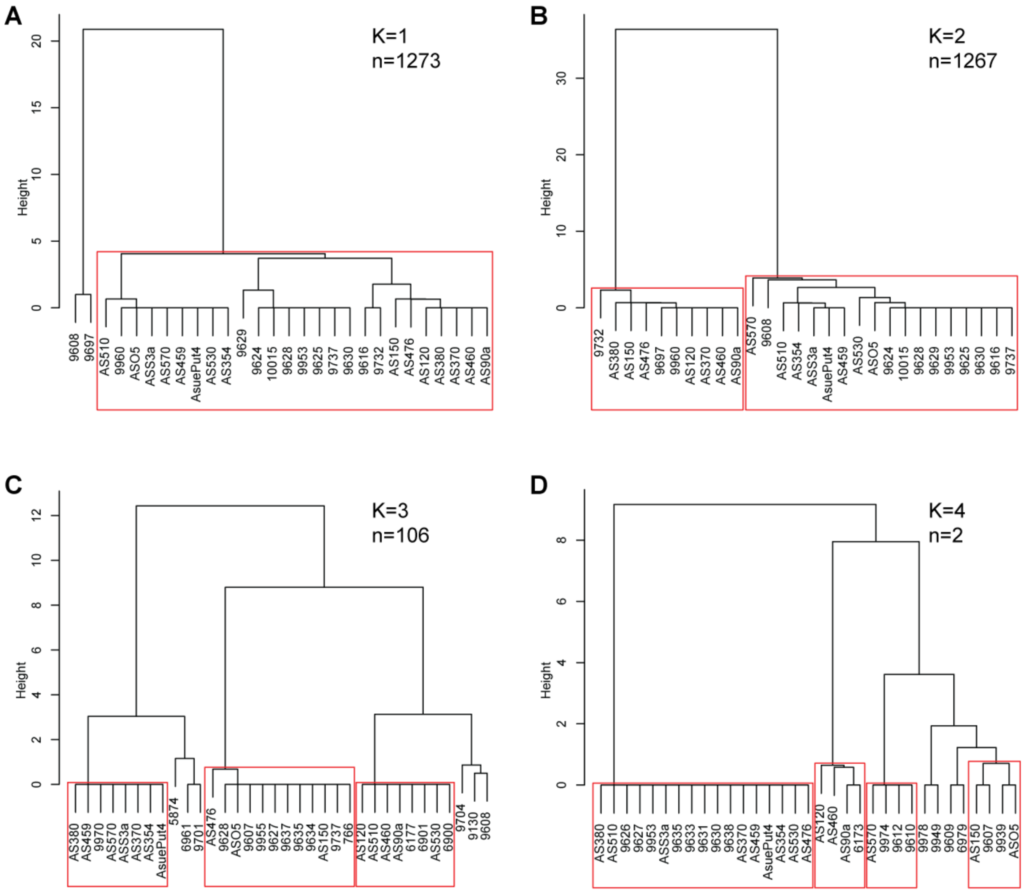


Fig. 4. Number of founder haplotypes in *A. suecica*. (**A-D**) Examples of the different number of founder haplotypes in *A. suecica*. *A. suecica* accessions are divided into clusters (red boxes) which also include accessions of *A. thaliana* and are strongly supported by p-values following multiscale bootstrap resampling (*Materials and Methods*).

Finally we considered the transition to selfing in *A. suecica*. *A. suecica* is self-compatible (Säll, Lind-Halldén et al. 2004), which agrees with the general association between polyploidy and selfing in plants (Barringer 2007). Selfing often evolves through the loss of the self-incompatibility system, which is controlled by the S-locus in flowering plants (Barrett 2002). In the genus *Arabidopsis*, the tightly linked male *SCR* (S-locus cysteine-rich protein - present in the pollen coat) and female *SRK* (S-locus receptor kinase - expressed on the surface of the stigma) determine the specificities of the self-recognition system: the male gene serving as the ligand and the female gene being a receptor (Takayama and Isogai 2005). Recognition of *SCR* by the *SRK* protein triggers a downstream signaling pathway that prevents pollen tube growth (Comai, Tyagi et al. 2000, Takayama and Isogai 2005, Chapman and Goring 2010). In the predominantly selfing *A. thaliana*, the S-locus is nonfunctional due to several loss-of-function mutations (Nasrallah, Liu et al. 2002, Liu, Sherman-Broyles et al. 2007, Sherman-Broyles, Boggs et al. 2007, Tang, Toomajian et al. 2007, Shimizu, Shimizu-Inatsugi et al. 2008, Boggs, Nasrallah et al. 2009, Tsuchimatsu, Suwabe et al. 2010)

"Synthetic *A. suecica*" F1 hybrids, produced by fertilizing colchicine-induced tetraploids of *A. thaliana* with pollen from naturally tetraploid *A. arenosa,* were not immediately selfing, and exhibited many abnormal phenotypes compared to natural *A. suecica* (Chen, Comai et al.

1998, Comai, Tyagi et al. 2000), however, they became increasing self-compatible after several rounds of forced self-pollination (Z.J. Chen, personal communication). However, a single *A. arenosa* collect (*Care-1*) was used in these crosses, and it is thus not known if other combinations of parents produce fully self-compatible hybrids.

In our sample of natural *A. suecica*, we found that the S-locus inherited from *A. thaliana* is fixed for the 213-bp inversion in the *SCR* gene (supplementary fig. S3) that is suggested to have led to loss of self-incompatibility in *A. thaliana* (Tsuchimatsu, Suwabe et al. 2010). Therefore, *A. thaliana* was almost certainly already self-compatible when it contributed to *A. suecica*, supporting the notion that the transition to selfing is more ancient (Bechsgaard, Castric et al. 2006, Tang et al, 2007, Hu et al. 2011).

A. arenosa is an obligate outcrosser (Säll, Lind-Halldén et al. 2004), hence *A. suecica* should have inherited fully functional S-alleles from *A. arenosa*, and these must somehow have been rendered nonfunctional or silenced. *Arabidopsis* S-alleles have a complex dominance hierarchy, determined by a small RNA regulatory network (Tarutani, Shiba et al. 2010, Durand, Meheust et al. 2014) in which small RNAs from some S-alleles can silence expression of *SCR* on other S-alleles. The S-locus is a classic example of a long-term balancing selection, with suppressed recombination between SRK and SCR leading to highly diverged S-allele shared across species (Mable, Schierup et al. 2003, Castric and Vekemans 2004, Mable, Beland et al. 2004, Castric, Bechsgaard et al. 2008, Llaurens, Billiard et al. 2008). In other words, a S-allele from *A. arenosa* may be more closely related to, for example, a S-alleles from *A. halleri*, than to other *A. arenosa* alleles. The *A. thaliana* S-allele found in *A. suecica* is predicted (based on cross-species alignments; see fig. 5 and *Materials and Methods*) to be an ortholog of the *A. halleri* S-haplogroup 4, whereas the *A. arenosa* S-allele found in *A. suecica* is orthologous to S-haplogroup 2. Based on crosses in *A. halleri*, the former has been shown to be dominant over the latter (Llaurens, Billiard et al. 2008), meaning that in *A. suecica* the S-allele at the locus inherited from *A. arenosa* is predicted to be transcriptionally silenced by the S-allele at the locus inherited from *A. thaliana*. Further investigation showed that the main players ensuring such a dominance mechanism (miRNA-producing loci and target sites) are most probably functional in *A. suecica* (*Materials and Methods,* supplementary fig. S8).

Moreover, mapping *A. suecica* to a complete BAC-sequence of *A. halleri* S-allele 2 (*Materials and Methods*) revealed that the S-locus inherited from *A. arenosa* is fixed for a single S-allele, which exhibited a frame-shift mutation in *SCR* that is predicted to lead to loss of function (supplementary fig. S9). It is therefore possible that *A. suecica* was immediately at least partly self-compatible due to the non-functional allele from *A. thaliana* being dominant over the functional allele from *A. arenosa*, and that the loss-of-function mutation was fixed later — it could either have been fixed by drift, or by selection. Independent of explanation, this mutation rendered the species fully self-compatible. In support of the former explanation, however, there is no sign of a selective sweep in the S-allele region from *A. arenosa* (supplementary fig. S10).

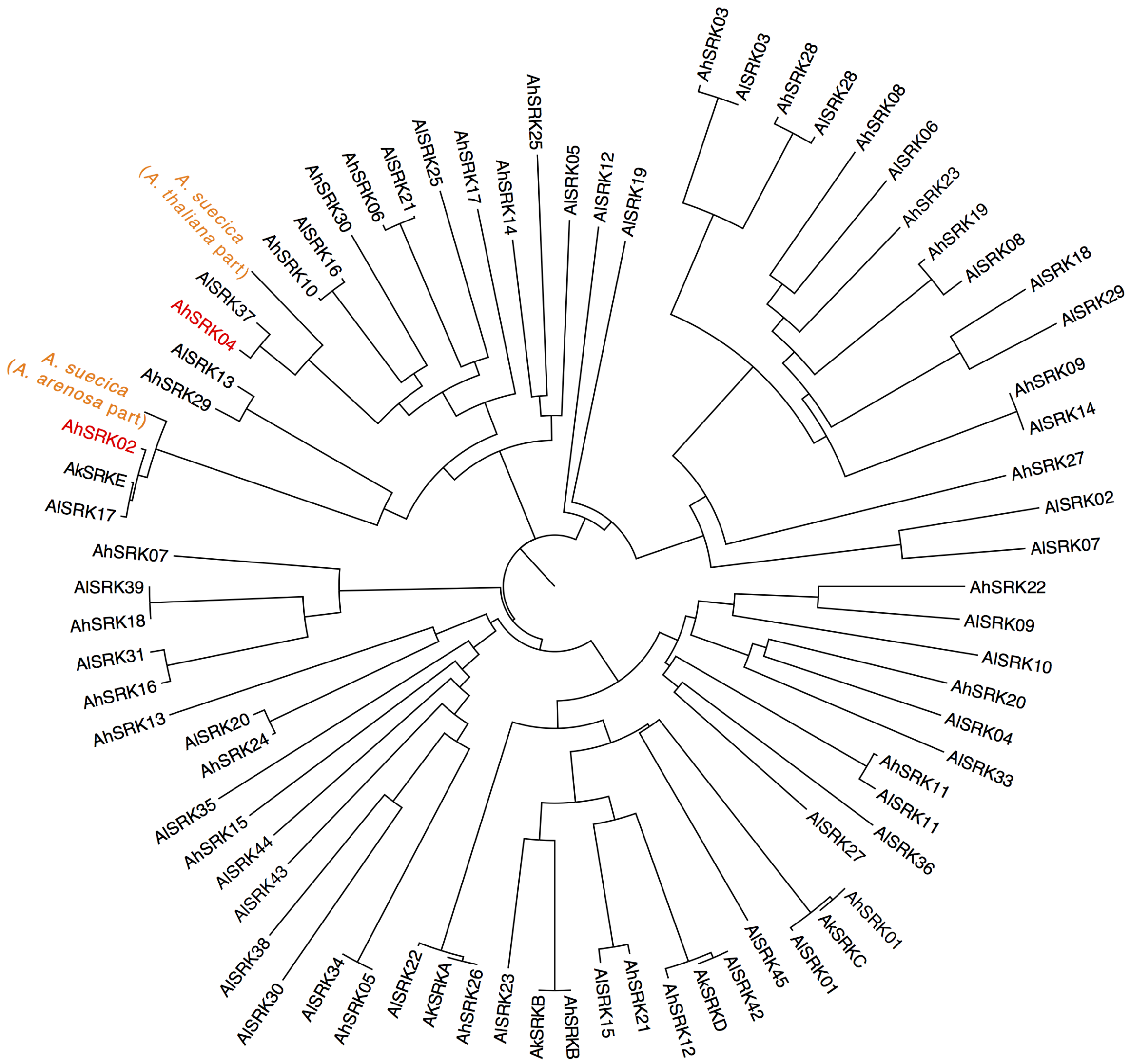


Fig. 5. The phylogeny of the *SRK* sequences belonging to *A. lyrata* (AlSRK), *A. halleri* (AhSRK), *A. kamchatica* (AkSRK) and *A. suecica*. The alignment is adapted from (Tsuchimatsu, Kaiser et al. 2012) and the phylogeny was generated using a neighbor-joining algorithm.

In conclusion, whole-genome sequencing of 15 natural *A. suecica* accessions from the entire species range revealed that the species was almost certainly founded by multiple hybridizations between *A. thaliana* (as the mother) and tetraploid *A. arenosa* (as the father), which is in line with generalizations made for the origins of the other well-established allotetraploid species from multiple founders (Soltis DE 2004, Shimizu-Inatsugi, Lihova et al. 2009, Vallejo-Marin, Buggs et al. 2015). Any scenario involving a single origin would have to invoke a completely outcrossed parent on the *A. thaliana* side (to explain genome-wide allele sharing), and subsequent gene flow between species with dramatically different karyotypes (to explain the regions of the genome that harbor more than two ancestral haplotypes). The species appears to have originated somewhere in central Eurasia in conjunction with the last glacial maximum, and subsequently migrated to Fennoscandinavia. A possibility is that hybridization was facilitated by the production of an unusually high frequency of unreduced *A. thaliana* ovules in a population containing both parental species, perhaps as a result of environmental stress (Madlung 2013, Vanneste, Maere et al. 2014, Mason and Pires 2015, Zhou, Mo et al. 2015). The transition to self-fertilization may have been facilitated by the dominance of the non-functional S-allele from *A. thaliana* over the S-allele from *A. arenosa*, followed by inactivation of the latter.

Materials and Methods

DNA extraction, library preparation and sequencing

Whole genomic DNA was extracted from fresh leaf material using the DNEasy Plant Mini Kit (Qiagen). Total genomic DNA libraries were prepared using a slightly modified Illumina Genomic DNA Sample Prep protocol. Briefly, 100-200 ng DNA was fragmented by sonication with Bioruptor (Diagenode); the peak of fragment sizes was about 400 bp. End-repair of sheared DNA fragments, A-tailing and adapter ligation were done with the NEXTflex^{TM} DNA Sequencing Kit (Bio Scientific). Adaptor-modified DNA was resolved on 1.5 % low melt agarose (Peqlab) gel. For size selection of the library, DNA was excised from the gel with the size range from 300 to 500bp. The paired-end DNA libraries were amplified for 10-12 cycles by PCR with VeraSeq PCR Mix (Biozym). Libraries were sequenced in 100-bp paired-end mode on Illumina HiSeq 2000 Analyzers using the manufacturer's standard cluster generation and sequencing protocols.

Raw sequencing data for 12 new *A. suecica* accessions were uploaded to NCBI SRA under BioProject ID PRJNA309929; data for 3 additional accessions were already published (Novikova, Hohmann et al. 2016) and are available under BioProject ID PRJNA284572 (supplementary table S1).

Read mapping and variants discovery

We mapped *A. suecica* reads to the *A. thaliana* (TAIR10) and the *A. lyrata* (v1.0) references simultaneously using the BWA-MEM algorithm from BWA (Li and Durbin 2009) (version 0.7.8) with an increased penalty of 15 for unpaired read pairs. The *A. lyrata* reference genome is the closest available reference to *A. arenosa*, and the successful mapping of *A. suecica* to a combination of the *A. thaliana* and *A. lyrata* genomes has been reported before (Henry, Dilkes et al. 2014). We used Samtools (Li, Handsaker et al. 2009) (version 0.1.19) to sort, index and remove potential duplications from the PCR amplification step of library preparation. We then performed a local realignment with IndelRealigner from Genome Analysis Toolkit (McKenna, Hanna et al. 2010, DePristo, Banks et al. 2011) (version 3.3.0). After filtering for uniquely aligned reads with Samtools, we called sites and variants using GATK UnifiedGenotyper with default parameters. We combined called sites from all *A. suecica* samples with CombineVariants from GATK and annotated sites using SNPeff (Cingolani, Platts et al. 2012). In order to decrease the number of variant calls from any misaligned homoeologous regions, we filtered out all the heterozygous calls, which were present in the population sample more than once.

We used previously published raw data from *A. arenosa* clade samples (45 individuals in total; see Novikova, Hohmann et al. 2016). *A. arenosa* reads were mapped to the *A. lyrata* (v1.0) reference, using the pipeline described above. All confident *A. arenosa* calls, including variants, were combined with the *A. suecica* calls for the *A. lyrata* component of the reference used for the mapping of *A. suecica*. The *A. thaliana*-derived portion of *A. suecica* was compared throughout the study with variant calls of Eurasian accessions taken from the 1001 *Arabidopsis* genome project (Consortium 2016).

Ancestral *A. thaliana* population(s) for *A. suecica*

Shared, private and differentially fixed polymorphism in *A. suecica* were compared with the parental species (fig. 1) and were calculated for all sites, for which data were available for at least 80% of individuals in both populations for each comparison. Polymorphism density in fig. 1 represents the number of polymorphic sites per aligned site in 200-kb windows along the genome. The R library 'circlize' was used to visualize data along the chromosomes of *A. suecica*.

Pairwise divergence between *A. thaliana* and *A. suecica* accessions was calculated with custom scripts as a percentage of diverged sites from all aligned sites, excluding indels. Fig. 2 shows minimal divergence between each *A. thaliana* accession and all *A. suecica* accessions. Maximum likelihood estimates of individual ancestries (with K=9 for the number of clusters) were done via ADMIXTURE (Alexander, Novembre et al. 2009), allowing for missing data. The R library 'ggmap' was used to visualize the data points on Google Maps. The R library 'raster' was used for visualizing the climate data from http://worldclim.org/paleo-climate with a resolution of 10 minutes.

In fig. 3B,C the average pairwise distance between 9629 and 6064 was calculated in 200-kb windows along the genome. The violet lines in fig. 3B,C show the position of the longest haplotype that is shared with *A. suecica* and is defined by using a threshold of a 0.1% error rate along the region of interest (indicated in orange).

Number of founder haplotypes in *A. suecica*

We divided the *A. thaliana* portion of the *A. suecica* genome into 200-kb windows, and estimated pairwise divergence between all *A. thaliana* and *A. suecica* accessions. We choose the 5 closest *A. thaliana* accessions for each *A. suecica* accession, which resulted in a range of 5 to 43 unique *A. thaliana* accessions, depending on the examined interval. For each interval, we calculated haplotype blocks and all possible haplotype phases for *A. suecica* together with the selected *A. thaliana* accessions, using PLINK (Purcell, Neale et al. 2007) with default parameters. Using consensus sequences of the most likely haplotype phases for each individual, we performed hierarchical clustering and calculated an uncertainty level for each cluster with the R package 'pvclust' (Suzuki and Shimodaira 2006) (method.hclust="ward.D2"). The uncertainty of each cluster was assessed using an 'AU' (Approximately Unbiased) p-value, which is computed by multiscale bootstrap resampling. In order to estimate the number of founder haplotypes in *A. suecica*, we counted, for each haplotype, the number of clusters with an AU greater than 99%, where *A. suecica* and *A. thaliana* accessions are present in the same cluster.

We ran coalescent simulations using the R library 'scrm' (analogous to ms (Hudson 2002)) under a linear growth model with a varying θ (measured in $4N_0$ generations) and population size at time t: $N_t=n*N_0$, where N_0 is the contemporary population size of *A. suecica* and N_t is the number of founders for *A. suecica*. These parameters were uniformly distributed (N_0 varied from 1000 to 100,000) and (N_t varied from 1 to 1000 for each N_0) and result in a total of 100,000 MS runs. For each MS run, 2648 gene genealogies were generated, where we calculated the number of ancestral lineages for time t = 16 Kya, which match our estimate for the origin time of *A. suecica* (supplementary fig. S7). A comparison with the observed number of ancestral lineages for 2648 loci (fig. 4) was conducted via a least squared distance method. An additional 1,000,000 simulations were run with a fixed N_0 of 5,000 and from this we chose 1000 simulations that were close to the observed data, allowing a posterior distribution of the N_t parameter to be inferred.

Dating the origin of *A. suecica*

Our estimation of the time of origin of *A. suecica* was based on the assumption that at loci inherited from one founder, the population diversity within *A. suecica* is generated solely by new mutations. Therefore, we can estimate the origin time of *A. suecica* simply as π/2 μ, where π is nucleotide diversity within *A. suecica* at the single founder region, μ is mutation rate and the generation time is one year (*A. suecica* is an annual plant). Here, we take the mutation rate of *A. suecica* to equal that of *A. thaliana*: 7 x 10(-9) base substitutions per site per generation (Ossowski, Schneeberger et al. 2010). We calculated the origin time of *A. suecica* as the expected coalescence time within accessions at loci which belong to the same founder haplotype: 95% CI [15.1 Kya, 16.6 Kya]. Confidence intervals for the median of the distribution were calculated using the basic bootstrap method in the R package 'boot'. We obtained a similar result for the estimated origin time of *A. suecica* (95% CI [14.1 Kya, 18.4 Kya]) by applying the same logic at the largest single founder region on the second chromosome of the *A. thaliana* portion of the *A. suecica* genome (between 7.8 and 16.2 Mbp). Nucleotide diversity was calculated in 200-kb windows along this region.

Coalescent rates and the scaled population size over time were inferred using MSMC (Schiffels and Durbin 2014) for 6 combinations of 4 randomly chosen *A. suecica* accessions separately for mapping to *A. thaliana* and *A. lyrata* (representing the *A. arenosa* portion of the *A. suecica* genome) chromosomes of the combined reference genome. Only intervals with a continuous coverage over 10 kb were chosen for the analysis.

Mechanism of self-compatibility in *A. suecica*

Combining de novo assembly and alignment tactics, we assigned S-haplogroups to *A. thaliana*- and *A. arenosa*-derived S-alleles in *A. suecica* using partial *SRK* sequences (see below). *A. thaliana*-derived and *A. arenosa*-derived S-alleles of *A. suecica* appear to be orthologous to corresponding *A. halleri* S-alleles 4 (AhS04) and 2 (AhS02) (fig. 5). AhS02 and AhS04 are members of the second most recessive class of *Arabidopsis* S-haplogroups (Durand, Meheust et al. 2014). However, the inferred pollen phenotype of *A. halleri* plants with a heterozygous AhS02/AhS04 genotype was that of AhS04 (Llaurens, Billiard et al. 2008). This suggests that the *A. thaliana*-derived S-haplogroup (an ortholog of AhS04) could be partially dominant over the *A. arenosa*-derived S-haplogroup (an ortholog of AhS02) in *A. suecica*; and that the *A. suecica* pollen phenotype should correspond to the *A. thaliana*-derived *SCR*, which is truncated and allows for selfing. Such a combination of S-haplogroups also appears to be present in all the analyzed *A. suecica* accessions and, most probably, fixed in the *A. suecica* species.

Mapping *A. suecica* to a complete BAC-sequence of *A. halleri* S-allele 2 (see below) revealed a loss-of-function mutation at the *SCR* gene (supplementary fig. S9) is fixed in *A. suecica*, while functionally and structurally important residues (supplementary fig. S11) are conserved in the *SRK* gene. It is therefore possible that *A. suecica* became a selfer following the frame-shift mutation in *SCR*. However, it is also possible that silencing of the *A. arenosa*-derived S-allele by the *A. thaliana*-derived S-allele provided an immediate selfing opportunity for *A. suecica*, followed by the subsequent pseudogenization of the *SCR* gene, making *A. suecica* an irrevocable selfer. In line with this hypothesis, we found that *mir867* expressed in the *A. thaliana* haplotype A of Col-0 as well as the corresponding *A. halleri* haplotype 4, which is predicted to be able to target the *A. halleri* haplotype 2 is fully identical to the mature miRNA-producing portion in *A. suecica* (supplementary fig. S8). Its target sequence in the orthologous *A. suecica SCR02* is also fully identical, suggesting that the silencing mechanism could have been active at the speciation time of *A. suecica*. Target sites of sRNA reads produced by *mir867* of AhS04 and the *A. thaliana* haplotype A were predicted by mapping to the AhSCR02 genomic sequence using a modified Smith-Waterman algorithm and a threshold of 18 (Durand, Meheust et al. 2014). sRNA sequencing data were from (Durand, Meheust et al. 2014) for Ah04 in *A. halleri* (GSM1378105) and from (Montgomery, Yoo et al. 2008) (ago1-25, GSE13605), (Mi, Cai et al. 2008) (AGO4-IP, GSE10036), and (Zheng, Ryvkin et al. 2010) (rdr6, GSE23439) for haplotype A in *A. thaliana*.

Assembly of the *A. suecica* accession ASS3a and the assignment of S-alleles

We assembled one *A. suecica* accession (ASS3a) that possessed the highest number of Illumina reads, using SOAPdenovo2 (Luo, Liu et al. 2012) (127mer version 1.4.10) with a kmer length equal to 73. We used all the reads, both at the contig and scaffold assembly level (asm_flags=3). The resulting assembly had an N50 length of scaffolds equal to 38,763 bp.

Using the *SCR* sequence from *A. thaliana* (Col-0) as a query (Shimizu, Shimizu-Inatsugi et al. 2008), we searched in our *de novo A. suecica* assembly for the scaffold containing *SCR*-like sequences using Blast (v. 2.2.28) and applying penalties for the opening and extension of gaps equal to 2 and 1, respectively. With this, we identified scaffold3258 as the scaffold that contains the *A. suecica SCR* sequence for the *A. thaliana*-derived portion of the genome.

We obtained available *SRK* sequences for *A. thaliana*, *A. halleri* and *A. lyrata* (Schierup, Mable et al. 2001, Charlesworth, Mable et al. 2003, Bechsgaard, Castric et al. 2006, Castric and Vekemans 2007, Tang, Toomajian et al. 2007, Castric, Bechsgaard et al. 2008, Castric, Bechsgaard et al. 2010, Tsuchimatsu, Kaiser et al. 2012). Using those *SRK* sequences as a query, we searched for the scaffolds containing *SRK*-like sequences in our *A. suecica* assembly using Blast with the same parameters. Combining the percent of identity with the bit score, we identified C3267705 and scaffold11240 as scaffolds containing the *SRK* sequences for the *A. thaliana* and *A. arenosa* portions of *A. suecica*, respectively. In order to assign the S-haplogroups in the ASS3a *A. suecica* accession, we incorporated the obtained *SRK* sequences from C3267705 and scaffold11240 scaffolds into the *SRK* alignment (Tsuchimatsu, Kaiser et al. 2012). We used CLC Main Workbench v7.0.2 (CLC bio, Aarhus, Denmark) to align the *SRK* sequences and applied default parameters for a ‘slow’ alignment: with gap open and extension cost being 10.0 and 1.0, respectively. A neighbor-joining tree from the *SRK* alignment was constructed using the same software, applying the Jukes-Cantor nucleotide distance as a measure.

In order to check whether all *A. suecica* accessions carry the same S-haplogroups, we included the partially assembled *A. arenosa* S-locus sequence from our *A. suecica* assembly (scaffold11240) to the combined reference genome of *A. thaliana* (TAIR10) and *A. lyrata* (v1.0) and mapped all the *A. suecica* accessions to this novel reference. Mapping was conducted using the pipeline described above, however, we did not filter for ‘primary’ aligned reads. The same pipeline was used for mapping of *A. suecica* accessions to the BAC-sequence of *A. halleri* S-allele 2, together with the *A. thaliana* and *A. lyrata* reference genomes. Consensus sequences of *SCR* and *SRK* genes were obtained with GATK FastaAlternateReferenceMaker (McKenna, Hanna et al. 2010, DePristo, Banks et al. 2011), aligned with MAFFT (Katoh and Standley 2013) (version 7) and visualized with JalView (Waterhouse, Procter et al. 2009).

Construction of BAC libraries

High Molecular Weight (HMW) DNA was prepared from young leaves of *Arabidopsis halleri* var. P21M53. For the extraction, 20 g of frozen leaf tissue was grounded to a powder in liquid nitrogen with a mortar and pestle in order to prepare megabase-size DNA embedded in agarose plugs. HMW DNA was prepared as described by Peterson *et al* (Peterson 2000) and modified as described by Gonthier and collaborators (Gonthier, Bellec et al. 2010). Embedded HMW DNA was partially digested with *Hind*III (New England Biolabs, Ipswich, Massachusetts), and subjected to two size selection steps by pulsed- field electrophoresis, using a BioRad CHEF Mapper system (Bio-Rad Laboratories, Hercules, California), and ligated to pIndigoBAC-5 *Hind*III-Cloning Ready vector (EpicentreBiotecnologies, Madison, Wisconsin). Pulsed-field migration programs, electrophoresis buffer, and ligation desalting conditions were performed according to Chalhoub *et al* (Chalhoub, Belcram et al. 2004). The BAC library is composed by 18 432 clones with a mean insert size of 110 kb and represents 6 genome equivalents.

PacBio RS II sequencing and assembly of the S-locus

2 µg of BAC clone Aha_P21M53_40F08 were pooled with 11 other BAC clones DNA to obtain a total amount of 24 µg. One library was generated using the standard Pacific Biosciences library preparation protocol for 8-12 kb libraries. This library was sequenced in one PacBio RS II SMRT Cell using the P4 polymerase in combination with the C2 chemistry (sequencing service following the standard operating procedures was provided by IGM Genomic Center).

Assembly of the PacBio RS II reads was performed following the HGAP workflow. The SMRT® Analysis (v2.2.0) software suite was used for HGAP implementation (https://github.com/PacificBiosciences/Bioinformatics-Training/wiki/HGAP). Reads were first aligned using BLASR ("Blasr on Pacific Biosciences repository website," n.d.; (Chaisson and Tesler 2012)) against "*Escherichia coli* str. K12 substr. DH10B, complete genome". Identified *E. coli* reads and low quality reads (read quality < 0.80 and read length < 500 bp) were removed from data used for the BAC clone sequences assembly. Vector sequences were trimmed as part of the assembly process. Each BAC assembly was individualized by matching its BES to the ends of assembled sequences using BLAST. Annotation of the *SRK* and *SCR* genes followed Goubet, Berges et al. (2012) and annotation of small RNA precursors followed Durand, Meheust et al. (2014).

Acknowledgments

This project received funding, in part. from the German Research Foundation DFG SPP 1529 to MN, the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (NOVEL project, grant agreement N°648321) to VC, a Grant-in-Aid for Scientific Research from the Japan Society for the Promotion of Science to TT. We thank Next Generation Sequencing Facility at Vienna Biocenter Core Facilities (VBCF) for obtaining the sequence data.

Supplementary Figures and Tables

Supplementary table S1. Analyzed *A. suecica* accessions.

accession	lat(N)	long(E)	Country	# reads	mapped reads after filtering, %	BioProject ID	BioSample_accession
AS120	61.54	16.33	Sweden	97,785,320	77.24	PRJNA309929	SAMN04442133
AS150	62.10	14.56	Sweden	176,097,226	78.40	PRJNA309929	SAMN04442134
AS354	63.43	17.27	Sweden	138,867,478	77.58	PRJNA309929	SAMN04442135
AS370	64.95	25.23	Finland	125,410,546	78.23	PRJNA309929	SAMN04442136
AS380	60.24	25.12	Finland	215,698,776	72.57	PRJNA309929	SAMN04442137
AS459	60.19	16.12	Sweden	316,754,654	78.55	PRJNA284572	SRS977003
AS460	60.48	15.48	Sweden	41,733,088	76.66	PRJNA309929	SAMN04442138
AS476	59.51	17.49	Sweden	90,375,686	79.24	PRJNA309929	SAMN04442139
AS510	60.11	24.59	Finland	76,672,068	77.08	PRJNA309929	SAMN04442140
AS530	61.15	24.20	Finland	47,870,552	76.90	PRJNA309929	SAMN04442141
AS570	60.46	16.57	Sweden	41,636,570	77.31	PRJNA309929	SAMN04442142
AS90a	63.47	17.05	Sweden	256,747,096	77.23	PRJNA309929	SAMN04442143
ASO5	63.70	18.25	Sweden	369,320,826	78.35	PRJNA284572	SRS977001
ASS3a	63.33	17.98	Sweden	418,718,928	73.58	PRJNA284572	SRS977000
AsuePut4	61.51	30.58	Russia	87,821,588	78.03	PRJNA309929	SAMN04442144

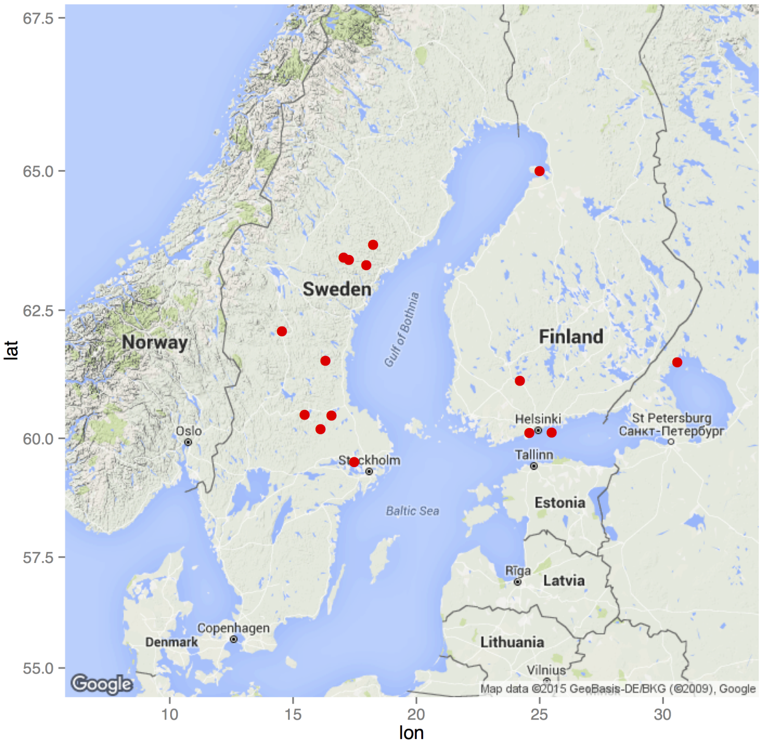


Fig. S1. Sampling locations of the 15 analyzed *A. suecica* accessions.

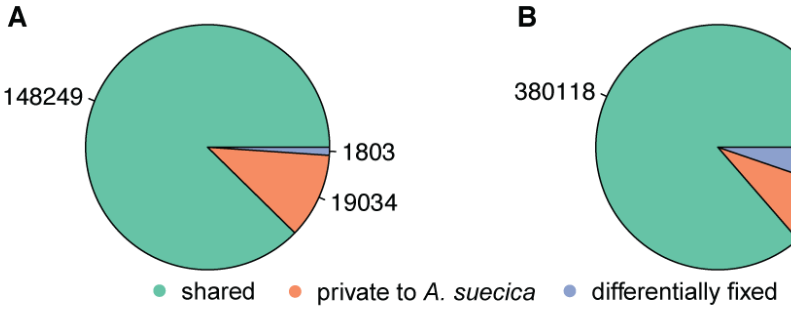


Fig. 2. The amount of shared SNPs (in green) between *A. suecica* and *A. thaliana* (**A**) and *A. arenosa* (**B**) contradicts the idea of a single hybridization event giving rise to *A. suecica*.

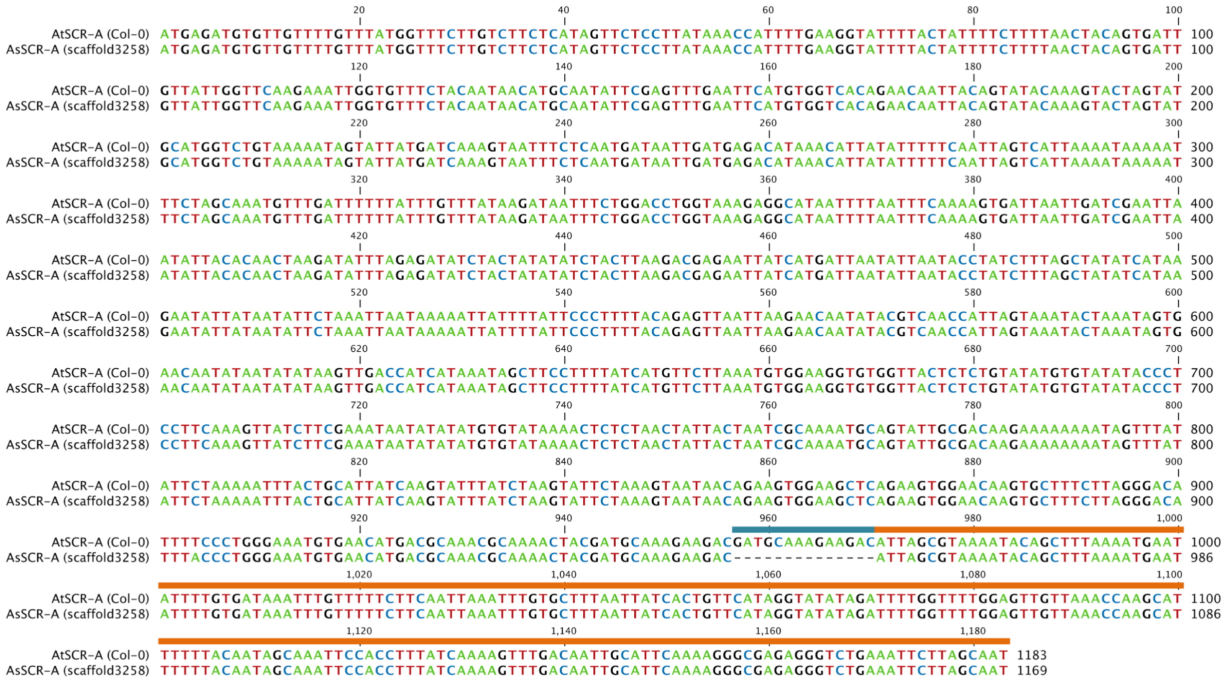


Fig. S3. Alignment of the *SCR-A* gene from *A. thaliana (Col-0*) and the *A .thaliana*-derived portion of *A. suecica* (scaffold3258). The *A. thaliana* sequence was obtained from the TAIR website (TAIR10 coordinate: Chr4.11382194-11383376; www.arabidopsis.org/). The alignment starts at the start codon of the *SCR-A* gene and ends with the gene-disruptive 213-bp inversion (Tsuchimatsu, Suwabe et al. 2010). The 213-bp inversion and the 14-bp duplication are highlighted the by orange and blue lines, respectively. Note that the 14-bp duplication is found in *Col-0* but is not widespread in *A. thaliana* accessions (Tsuchimatsu, Suwabe et al. 2010).

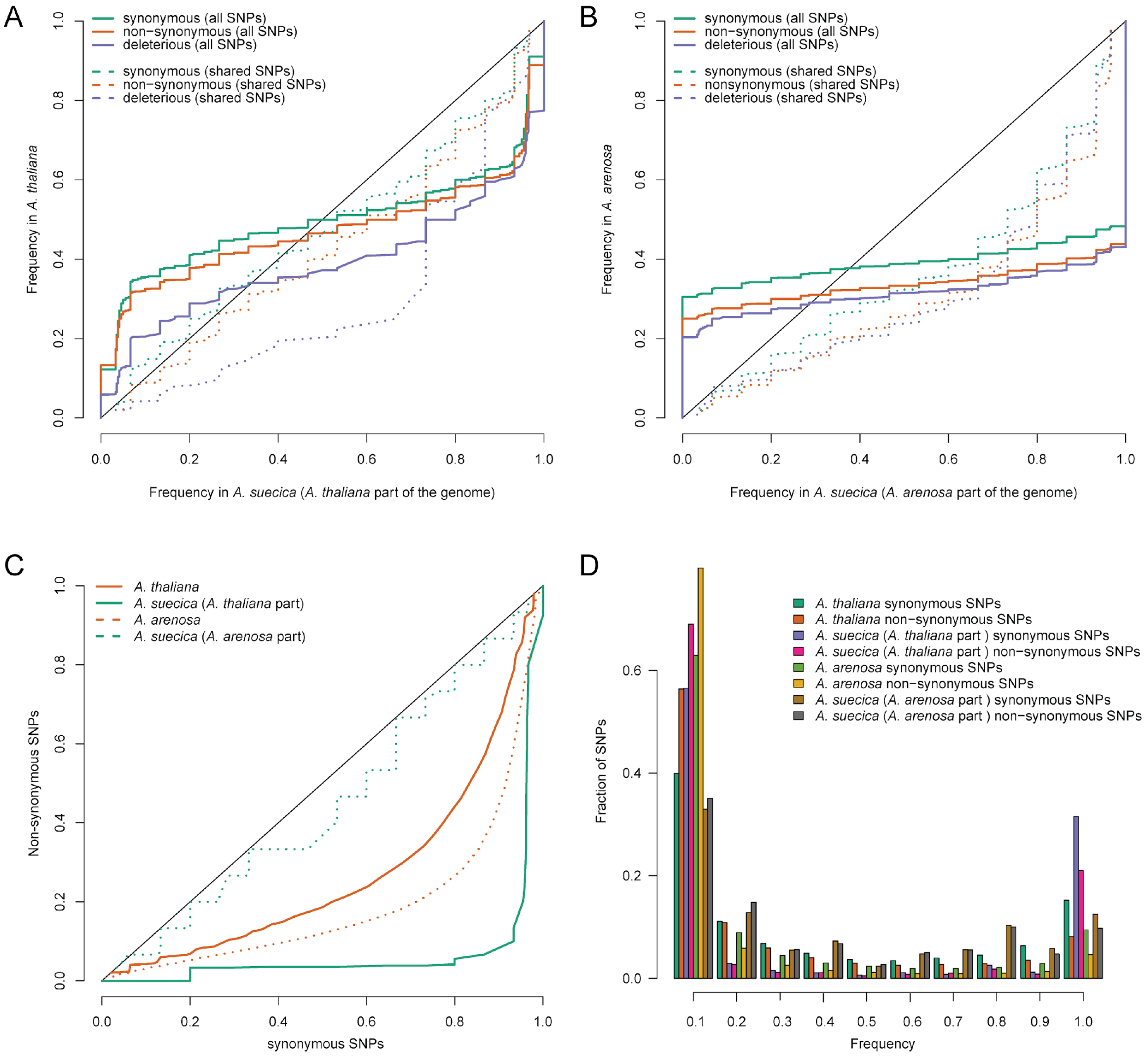


Fig. S4. The allele frequency distributions of putatively neutral and putatively deleterious polymorphisms in *A. suecica* and the parental species. (**A**-**B**) Comparison with the parental species demonstrates strong genetic drift presumably due to a bottleneck following hybridization. (**C**-**D**) Analysis of polymorphisms that are private to *A. suecica* demonstrate that purifying selection is acting in the hybrid species. “Deleterious” alleles were identified using SNPeff (Cingolani, Platts et al. 2012) and include stop codons and splice variants that have been gained/lost.

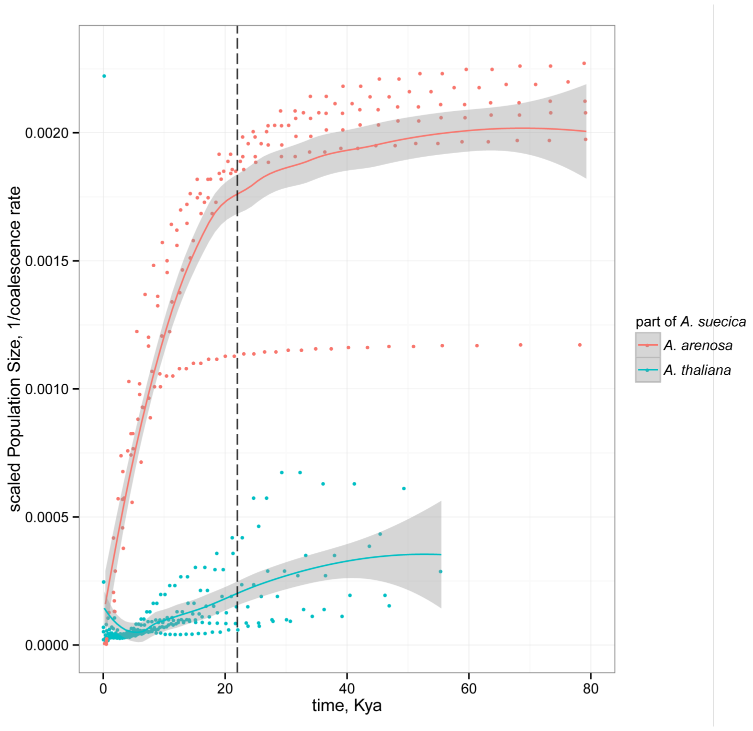


Fig. S5. Scaled population size changes over time inferred with MSMC (Schiffels and Durbin 2014) for the *A. thaliana* and *A. arenosa* portions of *A. suecica* (see *Materials and Methods*). Dots represent inferred population sizes at a particular time in the past for a given set of *A. suecica* accessions, while red and blue solid lines represent smoothed means with 0.95 confidence intervals (indicated by the grey area). Population size inferred from both portions of the genome decline following the last glacial maximum (black dashed line at 22 Kya) consistent with the bottleneck associated with *A. suecica's* origin.

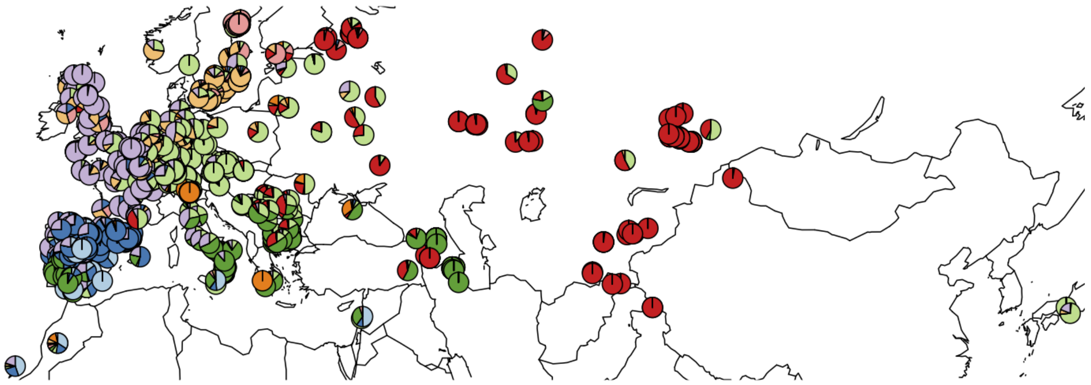


Fig. S6. The geographical distribution of the *A. thaliana* ancestral population (in red) for *A. suecica*, based on ADMIXTURE, with K=9 (Alexander, Novembre, and Lange 2009). All *A. suecica* accessions (not shown on this map) are assigned to the red cluster. *A. thaliana* accessions are depicted by pie charts that represent maximum likelihood assignments of individual ancestries, and these are marked by different colors.

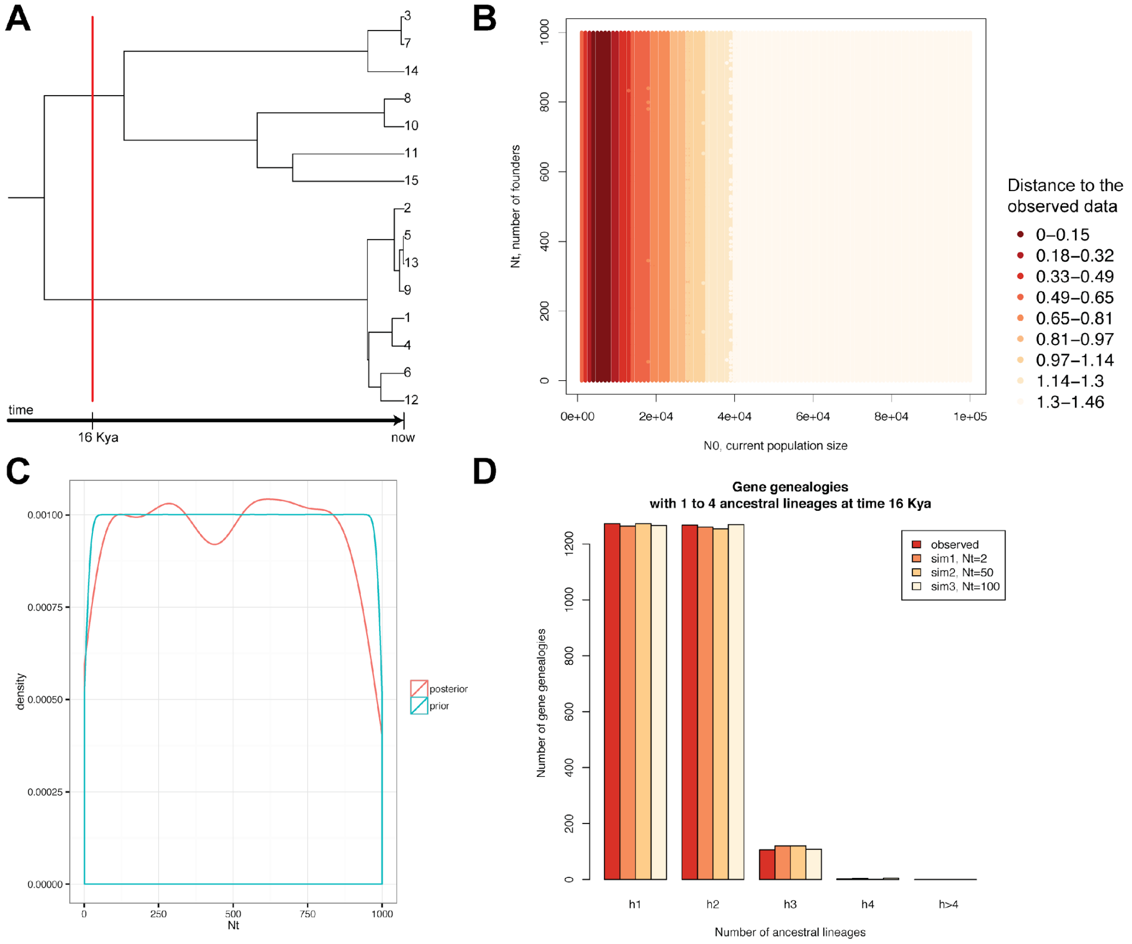


Fig. S7. The observed distribution of founder lineages is compatible with a wide range for the number of founders. (**A**) An example of a simulated gene genealogy with 2 ancestral lineages at the origin time of *A. suecica* - 16 Kya. (**B**) The distance to the observed data largely depends on the current population size (minimum distance at N_0=5,000), but not on the number of founders (ancestral lineages at time 16 Kya). (**C**) Prior and posterior distributions for the number of founders (N_t parameter with N_0 fixed at 5,000). The simulations that are closest to the observed data, with a 0.001 tolerance, were chosen to infer the posterior distribution of the N_t parameter. (**D**) The observed counts for the number of gene genealogies with a specific ancestral lineages at 16 Kya, compared with 3 of the simulations: varying the number of founders gives a similar count of gene genealogies.

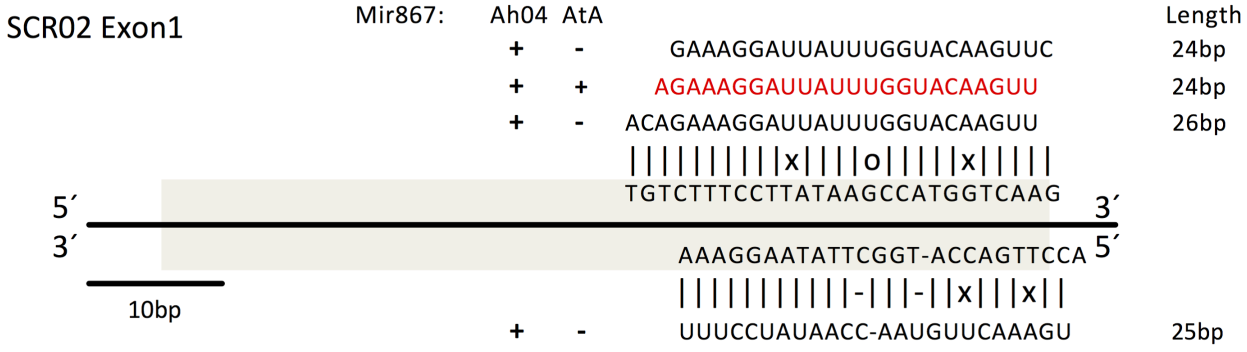

SCR02 Exon1
Mir867:
Ah04
AtA
Length
+
-
GAAAGGAUUAUUUGGUACAAGUUC
24bp
+
+
AGAAAGGAUUAUUUGGUACAAGUU
24bp
+
-
ACAGAAAGGAUUAUUUGGUACAAGUU
26bp
||||||||||x|||o||||x|||||
TGTCTTTCCTTATAAGCCATGGTCAAG
5´
3´
3´
5´
AAAGGAATATTCGGT-ACCAGTTCCA
|||||||||||-|||-||x|||x||
10bp
+
-
UUUCCUAUAACC-AAUGUUCAAAGU
25bp

Fig. S8. Mir867 expressed in the *A. halleri* S-locus haplotype 4 (Ah04 mir867) and the *A. thaliana* S-locus haplotype A (AtA mir867) is able to target the first exon of *A. halleri*/*A. suecica* SCR02 (denoted

by the grey box). Four different small RNAs from the S-locus Ah04 target SCR02, one of which is conserved with mir867 from AtA (highlighted in red).

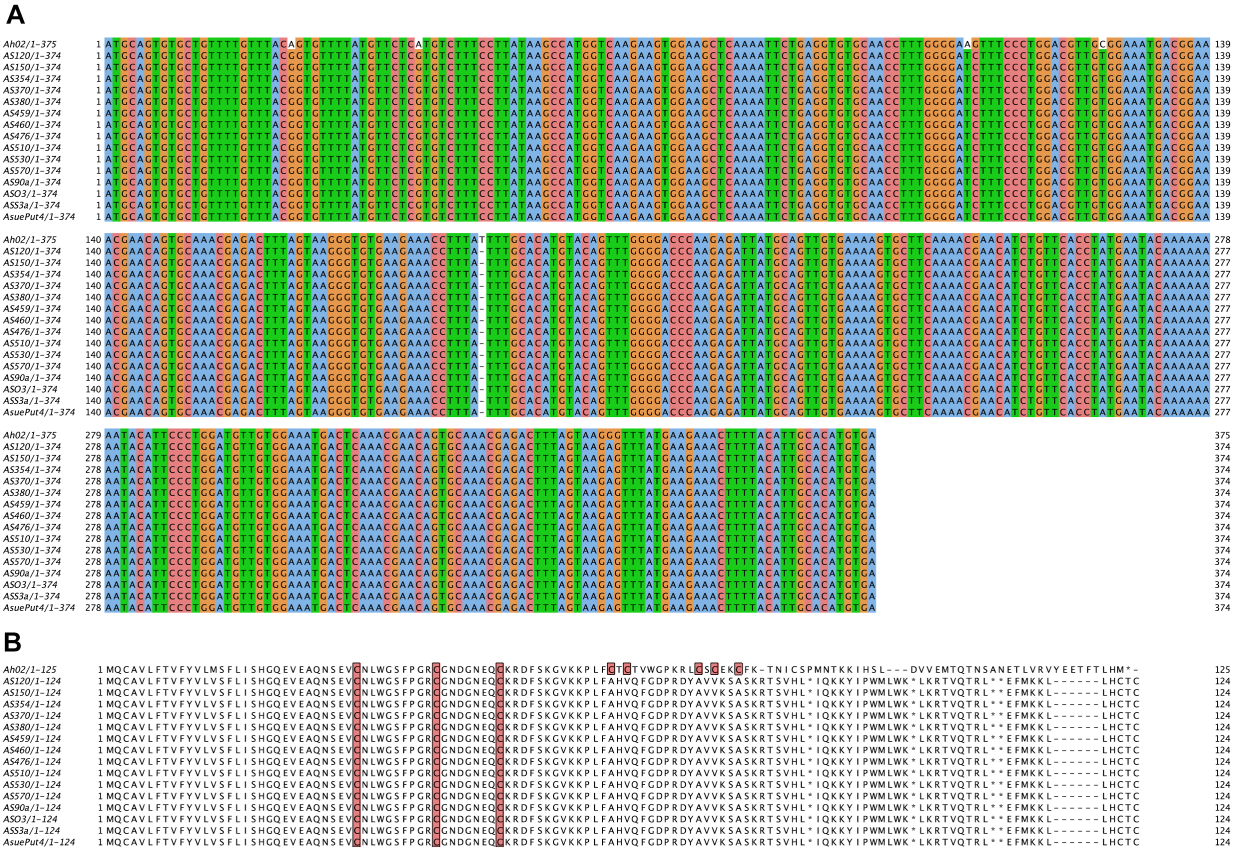


Fig. S9. Multiple sequence alignment of *SCR* from *A. halleri* (Ah02 S-haplogroup orthologous to *A. arenosa*-derived S-haplogroup in *A. suecica*) and *A. suecica* (consensus sequence from mapping to Ah02) at the DNA (**A**) and protein (**B**) level. The *A. arenosa*-derived *SCR* gene is likely to be non-functional in *A. suecica*, because it contains a frameshift mutation fixed in all *A. suecica* accessions compared to a functional *A. halleri SCR* gene in the orthologous S-haplogroup. The frameshift leads to the loss of 5 out of 8 conserved cysteines (indicated with red color on supplementary fig. S9B) that are important for protein structure (Mishima, Takayama et al. 2003, Tsuchimatsu, Suwabe et al. 2010).

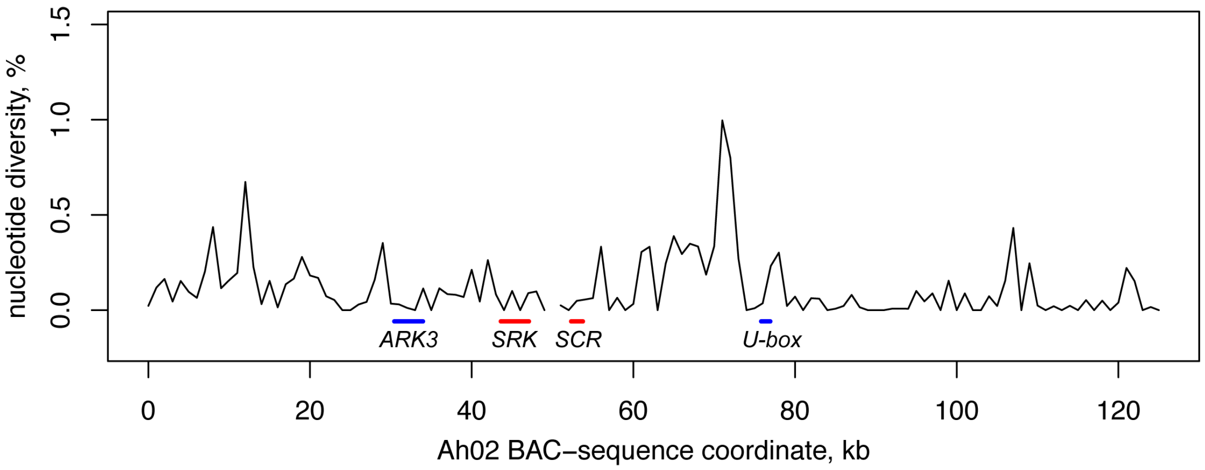


Fig. S10. *A. suecica* nucleotide diversity along Ah02 BAC-sequence, calculated in 1-kb windows.

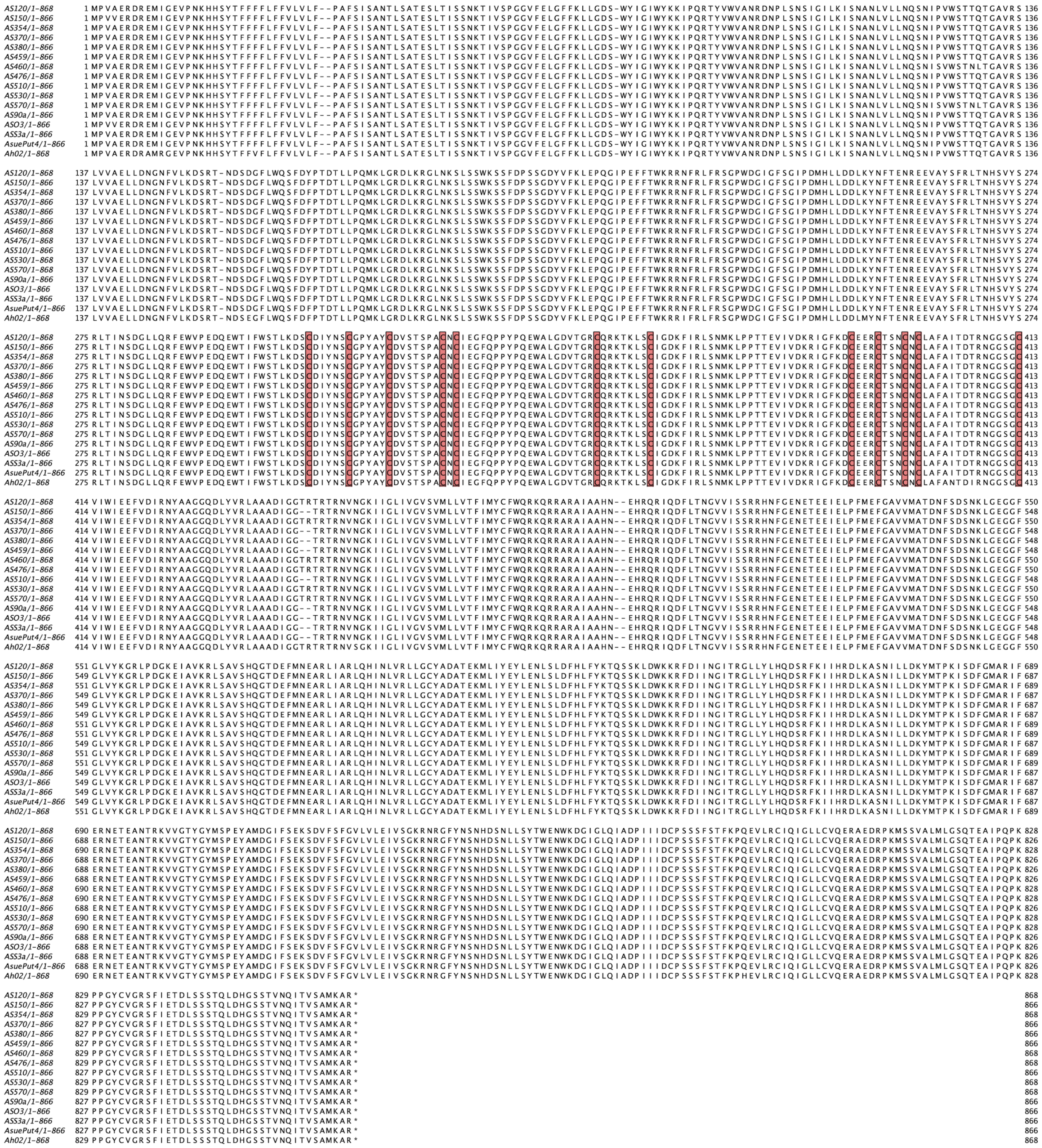


Fig. S11. Multiple alignment of predicted amino acid *SRK* sequence from *A. suecica* and *A. halleri* (Ah02 S-haplogroup). The red color indicates the 12 conserved cysteines that have been suggested to be important for the structure of the protein (Kusaba, Nishio et al. 1997, Naithani, Chookajorn et al. 2007, Tsuchimatsu, Kaiser et al. 2012).

References

Alexander, D. H., J. Novembre and K. Lange (2009). "Fast model-based estimation of ancestry in unrelated individuals." Genome Research **19**: 1655-1664.

Barrett, S. C. (2002). "The evolution of plant sexual diversity." Nat Rev Genet **3**(4): 274-284.

Barringer, B. C. (2007). "Polyploidy and self-fertilization in flowering plants." American Journal of Botany **94**: 1527-1533.

Bechsgaard, J. S., V. Castric, D. Charlesworth, X. Vekemans and M. H. Schierup (2006). "The transition to self-compatibility in Arabidopsis thaliana and evolution within S-haplotypes over 10 Myr." Molecular Biology and Evolution **23**: 1741-1750.

Boggs, N. A., J. B. Nasrallah and M. E. Nasrallah (2009). "Independent S-locus mutations caused self-fertility in Arabidopsis thaliana." PLoS genetics **5**: e1000426.

Castric, V., J. Bechsgaard, M. H. Schierup and X. Vekemans (2008). "Repeated Adaptive Introgression at a Gene under Multiallelic Balancing Selection." PLoS Genetics **4**: e1000168.

Castric, V., J. S. Bechsgaard, S. Grenier, R. Noureddine, M. H. Schierup and X. Vekemans (2010). "Molecular evolution within and between self-incompatibility specificities." Molecular Biology and Evolution **27**: 11-20.

Castric, V. and X. Vekemans (2004). "Plant self-incompatibility in natural populations: a critical assessment of recent theoretical and empirical advances." Molecular Ecology **13**: 2873-2889.

Castric, V. and X. Vekemans (2007). "Evolution under strong balancing selection: how many codons determine specificity at the female self-incompatibility gene SRK in Brassicaceae?" BMC evolutionary biology **7**: 132.

Chaisson, M. J. and G. Tesler (2012). "Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory." BMC Bioinformatics **13**: 238.

Chalhoub, B., H. Belcram and M. Caboche (2004). "Efficient cloning of plant genomes into bacterial artificial chromosome (BAC) libraries with larger and more uniform insert size." Plant Biotechnol J **2**(3): 181-188.

Chang, P. L., B. P. Dilkes, M. McMahon, L. Comai and S. V. Nuzhdin (2010). "Homoeolog-specific retention and use in allotetraploid Arabidopsis suecica depends on parent of origin and network partners." Genome Biology **11**: R125.

Chapman, L. A. and D. R. Goring (2010). "Pollen-pistil interactions regulating successful fertilization in the Brassicaceae." J Exp Bot **61**(7): 1987-1999.

Charlesworth, D., B. K. Mable, M. H. Schierup, C. Bartolomé and P. Awadalla (2003). "Diversity and linkage of genes in the self-incompatibility gene family in Arabidopsis lyrata." Genetics **164**: 1519-1535.

Chen, Z. J., L. Comai and C. S. Pikaard (1998). "Gene dosage and stochastic effects determine the severity and direction of uniparental ribosomal RNA gene silencing (nucleolar dominance) in Arabidopsis allopolyploids." Proc Natl Acad Sci U S A **95**(25): 14891-14896.

Cingolani, P., A. Platts, L. Wang le, M. Coon, T. Nguyen, L. Wang, S. J. Land, X. Lu and D. M. Ruden (2012). "A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3." Fly (Austin) **6**(2): 80-92.

Comai, L., A. P. Tyagi and M. A. Lysak (2003). "FISH analysis of meiosis in Arabidopsis allopolyploids." Chromosome Res **11**(3): 217-226.

Comai, L., A. P. Tyagi, K. Winter, R. Holmes-Davis, S. H. Reynolds, Y. Stevens and B. Byers (2000). "Phenotypic instability and rapid gene silencing in newly formed arabidopsis allotetraploids." Plant Cell **12**(9): 1551-1568.

Consortium, T. G. (2016). "1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana." Cell.

Costa-Nunes, P., O. Pontes, S. B. Preuss and C. S. Pikaard (2010). "Extra views on RNA-dependent DNA methylation and MBD6-dependent heterochromatin formation in nucleolar dominance." Nucleus (Austin, Tex.) **1**: 254-259.

DePristo, M. A., E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, M. Hanna, A. McKenna, T. J. Fennell, A. M. Kernytsky, A. Y. Sivachenko, K. Cibulskis, S. B. Gabriel, D. Altshuler and M. J. Daly (2011). "A framework for variation discovery and genotyping using next-generation DNA sequencing data." Nature Genetics **43**: 491-498.

Durand, E., R. Meheust, M. Soucaze, P. M. Goubet, S. Gallina, C. Poux, I. Fobis-Loisy, E. Guillon, T. Gaude, A. Sarazin, M. Figeac, E. Prat, W. Marande, H. Berges, X. Vekemans, S. Billiard and V. Castric (2014). "Dominance hierarchy arising from the evolution of a complex small RNA regulatory network." Science **346**(6214): 1200-1205.

Gonthier, L., A. Bellec, C. Blassiau, E. Prat, N. Helmstetter, C. Rambaud, B. Huss, T. Hendriks, H. Berges and M. C. Quillet (2010). "Construction and characterization of two BAC libraries representing a deep-coverage of the genome of chicory (Cichorium intybus L., Asteraceae)." BMC Res Notes **3**: 225.

Goubet, P. M., H. Berges, A. Bellec, E. Prat, N. Helmstetter, S. Mangenot, S. Gallina, A. C. Holl, I. Fobis-Loisy, X. Vekemans and V. Castric (2012). "Contrasted patterns of molecular evolution in dominant and recessive self-incompatibility haplotypes in Arabidopsis." PLoS Genet **8**(3): e1002495.

Griffiths, S., R. Sharp, T. N. Foote, I. Bertin, M. Wanous, S. Reader, I. Colas and G. Moore (2006). "Molecular characterization of Ph1 as a major chromosome pairing locus in polyploid wheat." Nature **439**(7077): 749-752.

Ha, M., J. Lu, L. Tian, V. Ramachandran, K. D. Kasschau, E. J. Chapman, J. C. Carrington, X. Chen, X.-J. Wang and Z. J. Chen (2009). "Small RNAs serve as a genetic buffer against genomic shock in Arabidopsis interspecific hybrids and allopolyploids." Proceedings of the National Academy of Sciences of the United States of America **106**: 17835-17840.

Henry, I. M., B. P. Dilkes, A. Tyagi, J. Gao, B. Christensen and L. Comai (2014). "The BOY NAMED SUE quantitative trait locus confers increased meiotic stability to an adapted natural allopolyploid of Arabidopsis." Plant Cell **26**(1): 181-194.

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G. and Jarvis, A. (2005). "Very high resolution interpolated climate surfaces for global land areas." Int. J. Climatol. **25**: 1965–1978.

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

Hudson, R. R. (2002). "Generating samples under a Wright-Fisher neutral model of genetic variation." Bioinformatics **18**(2): 337-338.

Hurka, K. M. a. (1995). "Allopolyploid Origin of Arabidopsis suecica (Fries) Norrlin: Evidence from Chloroplast and Nuclear Genome Markers." Botanica Acta **108**(5): 449–456.

Jakobsson, M., J. Hagenblad, S. Tavaré, T. Säll, C. Halldén, C. Lind-Halldén and M. Nordborg (2006). "A unique recent origin of the allotetraploid species Arabidopsis suecica: Evidence from nuclear DNA markers." Molecular Biology and Evolution **23**: 1217-1231.

Jiao, Y., N. J. Wickett, S. Ayyampalayam, A. S. Chanderbali, L. Landherr, P. E. Ralph, L. P. Tomsho, Y. Hu, H. Liang, P. S. Soltis, D. E. Soltis, S. W. Clifton, S. E. Schlarbaum, S. C. Schuster, H. Ma, J. Leebens-Mack and C. W. dePamphilis (2011). "Ancestral polyploidy in seed plants and angiosperms." Nature **473**(7345): 97-100.

Katoh, K. and D. M. Standley (2013). "MAFFT multiple sequence alignment software version 7: improvements in performance and usability." Mol Biol Evol **30**(4): 772-780.

Kusaba, M., T. Nishio, Y. Satta, K. Hinata and D. Ockendon (1997). "Striking sequence similarity in inter- and intra-specific comparisons of class I SLG alleles from Brassica oleracea and Brassica campestris: implications for the evolution and recognition mechanism." Proc Natl Acad Sci U S A **94**(14): 7673-7678.

Li, H. and R. Durbin (2009). "Fast and accurate short read alignment with Burrows-Wheeler transform." Bioinformatics (Oxford, England) **25**: 1754-1760.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin and G. P. D. P. Subgroup (2009). "The Sequence Alignment/Map format and SAMtools." Bioinformatics (Oxford, England) **25**: 2078-2079.

Liu, P., S. Sherman-Broyles, M. E. Nasrallah and J. B. Nasrallah (2007). "A cryptic modifier causing transient self-incompatibility in Arabidopsis thaliana." Curr Biol **17**(8): 734-740.

Llaurens, V., S. Billiard, J.-B. Leducq, V. Castric, E. K. Klein and X. Vekemans (2008). "Does frequency-dependent selection with complex dominance interactions accurately predict allelic frequencies at the self-incompatibility locus in Arabidopsis halleri?" Evolution; International Journal of Organic Evolution **62**: 2545-2557.

Long, Q., F. A. Rabanal, D. Meng, C. D. Huber, A. Farlow, A. Platzer, Q. Zhang, B. J. Vilhjalmsson, A. Korte, V. Nizhynska, V. Voronin, P. Korte, L. Sedman, T. Mandakova, M. A. Lysak, U. Seren, I. Hellmann and M. Nordborg (2013). "Massive genomic variation and strong selection in Arabidopsis thaliana lines from Sweden." Nat Genet **45**(8): 884-890.

Luo, R., B. Liu, Y. Xie, Z. Li, W. Huang, J. Yuan, G. He, Y. Chen, Q. Pan, Y. Liu, J. Tang, G. Wu, H. Zhang, Y. Shi, Y. Liu, C. Yu, B. Wang, Y. Lu, C. Han, D. W. Cheung, S. M. Yiu, S. Peng, Z. Xiaoqian, G. Liu, X. Liao, Y. Li, H. Yang, J. Wang, T. W. Lam and J. Wang (2012). "SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler." Gigascience **1**(1): 18.

Mable, B. K., J. Beland and C. Di Berardo (2004). "Inheritance and dominance of self-incompatibility alleles in polyploid Arabidopsis lyrata." Heredity **93**: 476-486.

Mable, B. K., M. H. Schierup and D. Charlesworth (2003). "Estimating the number, frequency, and dominance of S-alleles in a natural population of Arabidopsis lyrata(Brassicaceae) with sporophytic control of self-incompatibility." Heredity **90**: 422-431.

Madlung, A. (2013). "Polyploidy and its effect on evolutionary success: old questions revisited with new tools." Heredity (Edinb) **110**(2): 99-104.

Madlung, A., A. P. Tyagi, B. Watson, H. Jiang, T. Kagochi, R. W. Doerge, R. Martienssen and L. Comai (2005). "Genomic changes in synthetic Arabidopsis polyploids." Plant J **41**(2): 221-230.

Mason, A. S. and J. C. Pires (2015). "Unreduced gametes: meiotic mishap or evolutionary mechanism?" Trends Genet **31**(1): 5-10.

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

Mi, S., T. Cai, Y. Hu, Y. Chen, E. Hodges, F. Ni, L. Wu, S. Li, H. Zhou, C. Long, S. Chen, G. J. Hannon and Y. Qi (2008). "Sorting of small RNAs into Arabidopsis argonaute complexes is directed by the 5' terminal nucleotide." Cell **133**(1): 116-127.

Miller, M., Q. Song, X. Shi, T. E. Juenger and Z. J. Chen (2015). "Natural variation in timing of stress-responsive gene expression predicts heterosis in intraspecific hybrids of Arabidopsis." Nat Commun **6**: 7453.

Mishima, M., S. Takayama, K. Sasaki, J. G. Jee, C. Kojima, A. Isogai and M. Shirakawa (2003). "Structure of the male determinant factor for Brassica self-incompatibility." J Biol Chem **278**(38): 36389-36395.

Montgomery, T. A., S. J. Yoo, N. Fahlgren, S. D. Gilbert, M. D. Howell, C. M. Sullivan, A. Alexander, G. Nguyen, E. Allen, J. H. Ahn and J. C. Carrington (2008). "AGO1-miR173 complex initiates phased siRNA formation in plants." Proc Natl Acad Sci U S A **105**(51): 20055-20062.

N., H. (1957). "Cardaminopsis suecica (Fr.) Hiit., A northern amphidiploid species." Bull. Jard. Bot. État. Bruxelles. **271**(1): 591–604.

Naithani, S., T. Chookajorn, D. R. Ripoll and J. B. Nasrallah (2007). "Structural modules for receptor dimerization in the S-locus receptor kinase extracellular domain." Proc Natl Acad Sci U S A **104**(29): 12211-12216.

Nasrallah, M. E., P. Liu and J. B. Nasrallah (2002). "Generation of self-incompatible Arabidopsis thaliana by transfer of two S locus genes from A. lyrata." Science **297**(5579): 247-249.

Ng, D. W., M. Miller, H. H. Yu, T. Y. Huang, E. D. Kim, J. Lu, Q. Xie, C. R. McClung and Z. J. Chen (2014). "A Role for CHH Methylation in the Parent-of-Origin Effect on Altered Circadian Rhythms and Biomass Heterosis in Arabidopsis Intraspecific Hybrids." Plant Cell **26**(6): 2430-2440.

Ng, D. W., X. Shi, G. Nah and Z. J. Chen (2014). "High-throughput RNA-seq for allelic or locus-specific expression analysis in Arabidopsis-related species, hybrids, and allotetraploids." Methods Mol Biol **1112**: 33-48.

Nordborg, M. (1998). "On the probability of Neanderthal ancestry." Am J Hum Genet **63**(4): 1237-1240.

Novikova, P. Y., N. Hohmann, V. Nizhynska, T. Tsuchimatsu, J. Ali, G. Muir, A. Guggisberg, T. Paape, K. Schmid, O. M. Fedorenko, S. Holm, T. Sall, C. Schlotterer, K. Marhold, A. Widmer, J. Sese, K. K. Shimizu, D. Weigel, U. Kramer, M. A. Koch and M. Nordborg (2016). "Sequencing of the genus Arabidopsis identifies a complex history of nonbifurcating speciation and abundant trans-specific polymorphism." Nat Genet.

O'Kane, S. L., B. A. Schaal and I. A. Al-Shehbaz (1996). "The Origins of Arabidopsis suecica (Brassicaceae) as Indicated by Nuclear rDNA Sequences." Systematic Botany **21**: 559.

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

Pavelkova Ricankova, V., J. Robovsky and J. Riegert (2014). "Ecological structure of recent and last glacial mammalian faunas in northern Eurasia: the case of Altai-Sayan refugium." PLoS One **9**(1): e85056.

Peterson, D. G., Tomkins, J. P., Frisch, D. A., Wing, R. A., & Paterson, A. H. (2000). "Construction of Plant Bacterial Artificial Chromosome (BAC) Libraries: An Illustrated Guide." Journal of Agricultural Genomics 5: http://www.ncgr.org/research/jag.

Pikaard, C. S. (1999). "Nucleolar dominance and silencing of transcription." Trends in Plant Science **4**: 478-483.

Pontes, O., R. J. Lawrence, M. Silva, S. Preuss, P. Costa-Nunes, K. Earley, N. Neves, W. Viegas and C. S. Pikaard (2007). "Postembryonic establishment of megabase-scale gene silencing in nucleolar dominance." PLoS One **2**(11): e1157.

Pontvianne, F., T. Blevins, C. Chandrasekhara, W. Feng, H. Stroud, S. E. Jacobsen, S. D. Michaels and C. S. Pikaard (2012). "Histone methyltransferases regulating rRNA gene dose and dosage control in Arabidopsis." Genes & Development **26**: 945-957.

Price, R. A., I. A. Al-Shebaz, and J. D. Palmer (1994). "Systematic relationships of Arabidopsis: a molecular and morphological perspective in Arabidopsis." E. Meyerowitz and C. Somerville, eds. Cold Spring Harbour Laboratory Press, New York.: 7-19.

Prufer, K., F. Racimo, N. Patterson, F. Jay, S. Sankararaman, S. Sawyer, A. Heinze, G. Renaud, P. H. Sudmant, C. de Filippo, H. Li, S. Mallick, M. Dannemann, Q. Fu, M. Kircher, M. Kuhlwilm, M. Lachmann, M. Meyer, M. Ongyerth, M. Siebauer, C. Theunert, A. Tandon, P. Moorjani, J. Pickrell, J. C. Mullikin, S. H. Vohr, R. E. Green, I. Hellmann, P. L. Johnson, H. Blanche, H. Cann, J. O. Kitzman, J. Shendure, E. E. Eichler, E. S. Lein, T. E. Bakken, L. V. Golovanova, V. B. Doronichev, M. V. Shunkov, A. P. Derevianko, B. Viola, M. Slatkin, D. Reich, J. Kelso and S. Paabo (2014). "The complete genome sequence of a Neanderthal from the Altai Mountains." Nature **505**(7481): 43-49.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly and P. C. Sham (2007). "PLINK: a tool set for whole-genome association and population-based linkage analyses." American Journal of Human Genetics **81**: 559-575.

Ramsey, J. S., DW (1998). "Pathways, mechanisms, and rates of polyploid formation in flowering plants." ANNUAL REVIEW OF ECOLOGY AND SYSTEMATICS **29**: 467-501.

Reich, D., R. E. Green, M. Kircher, J. Krause, N. Patterson, E. Y. Durand, B. Viola, A. W. Briggs, U. Stenzel, P. L. Johnson, T. Maricic, J. M. Good, T. Marques-Bonet, C. Alkan, Q. Fu, S. Mallick, H. Li, M. Meyer, E. E. Eichler, M. Stoneking, M. Richards, S. Talamo, M. V. Shunkov, A. P. Derevianko, J. J. Hublin, J. Kelso, M. Slatkin and S. Paabo (2010). "Genetic history of an archaic hominin group from Denisova Cave in Siberia." Nature **468**(7327): 1053-1060.

Sall, T., M. Jakobsson, C. Lind-Hallden and C. Hallden (2003). "Chloroplast DNA indicates a single origin of the allotetraploid Arabidopsis suecica." J Evol Biol **16**(5): 1019-1029.

Säll, T., C. Lind-Halldén, M. Jakobsson and C. Halldén (2004). "Mode of reproduction in Arabidopsis suecica." Hereditas **141**: 313-317.

Schierup, M. H., B. K. Mable, P. Awadalla and D. Charlesworth (2001). "Identification and characterization of a polymorphic receptor kinase gene linked to the self-incompatibility locus of Arabidopsis lyrata." Genetics **158**: 387-399.

Schiffels, S. and R. Durbin (2014). "Inferring human population size and separation history from multiple genome sequences." Nature Genetics **46**: 919-925.

Sherman-Broyles, S., N. Boggs, A. Farkas, P. Liu, J. Vrebalov, M. E. Nasrallah and J. B. Nasrallah (2007). "S locus genes and the evolution of self-fertility in Arabidopsis thaliana." Plant Cell **19**(1): 94-106.

Shimizu, K. K., S. Fujii, K. Marhold, K. Watanabe and H. Kudoh (2005). "Arabidopsis kamchatica (Fisch. ex DC.) K. Shimizu & Kudoh and A. kamchatica subsp. kawasakiana (Makino) K. Shimizu & Kudoh, New Combinations." Acta phytotaxonomica et geobotanica **56**: 163–172.

Shimizu, K. K., R. Shimizu-Inatsugi, T. Tsuchimatsu and M. D. Purugganan (2008). "Independent origins of self-compatibility in Arabidopsis thaliana." Molecular Ecology **17**: 704-714.

Shimizu-Inatsugi, R., J. Lihova, H. Iwanaga, H. Kudoh, K. Marhold, O. Savolainen, K. Watanabe, V. V. Yakubov and K. K. Shimizu (2009). "The allopolyploid Arabidopsis kamchatica originated from multiple individuals of Arabidopsis lyrata and Arabidopsis halleri." Mol Ecol **18**(19): 4024-4048.

Solhaug, E. M., J. Ihinger, M. Jost, V. Gamboa, B. Marchant, D. Bradford, R. W. Doerge, A. Tyagi, A. Replogle and A. Madlung (2016). "Environmental Regulation of Heterosis in the Allopolyploid Arabidopsis suecica." Plant Physiol **170**(4): 2251-2263.

Soltis DE, S. P., Pires JC, Kovarik A, Tate JA, Mavrodiev E (2004). "Recent and recurrent polyploidy in Tragopogon (Asteraceae): cytogenetic, genomic and genetic comparisons." Biol J Linn Soc(82): 485–501.

Soltis, P. S., D. B. Marchant, Y. Van de Peer and D. E. Soltis (2015). "Polyploidy and genome evolution in plants." Curr Opin Genet Dev **35**: 119-125.

Suzuki, R. and H. Shimodaira (2006). "Pvclust: an R package for assessing the uncertainty in hierarchical clustering." Bioinformatics **22**: 1540-1542.

Takayama, S. and A. Isogai (2005). "Self-incompatibility in plants." Annu Rev Plant Biol **56**: 467-489.

Tang, C., C. Toomajian, S. Sherman-Broyles, V. Plagnol, Y.-L. Guo, T. T. Hu, R. M. Clark, J. B. Nasrallah, D. Weigel and M. Nordborg (2007). "The evolution of selfing in Arabidopsis thaliana." Science (New York, N.Y.) **317**: 1070-1072.

Tarasov, P. E., Volkova, V. S., Webb, T., Guiot, J., Andreev, A. A., Bezusko, L. G., Bezusko, T. V., Bykova, G. V., Dorofeyuk, N. I., Kvavadze, E. V., Osipova, I. M., Panova, N. K. and Sevastyanov, D. V. (2000). "Last glacial maximum biomes reconstructed from pollen and plant macrofossil data from northern Eurasia." Journal of Biogeography **27**: 609–620.

Tarutani, Y., H. Shiba, M. Iwano, T. Kakizaki, G. Suzuki, M. Watanabe, A. Isogai and S. Takayama (2010). "Trans-acting small RNA determines dominance relationships in Brassica self-incompatibility." Nature **466**(7309): 983-986.

Tian, L., X. Li, M. Ha, C. Zhang and Z. J. Chen (2014). "Genetic and epigenetic changes in a genomic region containing MIR172 in Arabidopsis allopolyploids and their progenitors." Heredity (Edinb) **112**(2): 207-214.

Tsuchimatsu, T., P. Kaiser, C.-L. Yew, J. B. Bachelier and K. K. Shimizu (2012). "Recent Loss of Self-Incompatibility by Degradation of the Male Component in Allotetraploid Arabidopsis kamchatica." PLoS Genetics **8**: e1002838.

Tsuchimatsu, T., K. Suwabe, R. Shimizu-Inatsugi, S. Isokawa, P. Pavlidis, T. Städler, G. Suzuki, S. Takayama, M. Watanabe and K. K. Shimizu (2010). "Evolution of self-compatibility in Arabidopsis by a mutation in the male specificity gene." Nature **464**: 1342-1346.

Vallejo-Marin, M., R. J. Buggs, A. M. Cooley and J. R. Puzey (2015). "Speciation by genome duplication: Repeated origins and genomic composition of the recently formed allopolyploid species Mimulus peregrinus." Evolution **69**(6): 1487-1500.

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

Vanneste, K., S. Maere and Y. Van de Peer (2014). "Tangled up in two: a burst of genome duplications at the end of the Cretaceous and the consequences for plant evolution." Philos Trans R Soc Lond B Biol Sci **369**(1648).

Vision, T. J., D. G. Brown and S. D. Tanksley (2000). "The origins of genomic duplications in Arabidopsis." Science **290**(5499): 2114-2117.

Wang, J., L. Tian, H. S. Lee, N. E. Wei, H. Jiang, B. Watson, A. Madlung, T. C. Osborn, R. W. Doerge, L. Comai and Z. J. Chen (2006). "Genomewide nonadditive gene regulation in Arabidopsis allotetraploids." Genetics **172**(1): 507-517.

Waterhouse, A. M., J. B. Procter, D. M. Martin, M. Clamp and G. J. Barton (2009). "Jalview Version 2--a multiple sequence alignment editor and analysis workbench." Bioinformatics **25**(9): 1189-1191.

Zheng, Q., P. Ryvkin, F. Li, I. Dragomir, O. Valladares, J. Yang, K. Cao, L. S. Wang and B. D. Gregory (2010). "Genome-wide double-stranded RNA sequencing reveals the functional significance of base-paired RNAs in Arabidopsis." PLoS Genet **6**(9): e1001141.

Zhou, X., X. Mo, M. Gui, X. Wu, Y. Jiang, L. Ma, Z. Shi, Y. Luo and W. Tang (2015). "Cytological, molecular mechanisms and temperature stress regulating production of diploid male gametes in Dianthus caryophyllus L." Plant Physiol Biochem **97**: 255-263.

Chapter2

This chapter is currently under revision at Nature Ecology & Evolution and was submitted on the 11th September 2020. This chapter is also available on bioRxiv with doi: https://doi.org/10.1101/2020.08.24.264432

Gradual evolution of allopolyploidy in *Arabidopsis suecica*

Robin Burns[1], Terezie Mandáková[2], Joanna Gunis[1], Luz Mayela Soto-Jiménez[1], Chang Liu[3], Martin A. Lysak[2], Polina Yu. Novikova[4,5*] and Magnus Nordborg[1*]

[1]Gregor Mendel Institute, Austrian Academy of Sciences, Vienna BioCenter, Vienna, Austria.
[2]CEITEC - Central European Institute of Technology, and Faculty of Science, Masaryk University, Brno, Czech Republic.
[3]Institute of Biology, University of Hohenheim, Garbenstrasse 30, 70599 Stuttgart, Germany.
[4]VIB-UGent Center for Plant Systems Biology, Ghent, Belgium.
[5]Department of Chromosome Biology, Max Planck Institute for Plant Breeding Research, Cologne, Germany.

***Corresponding authors: pnovikova@mpipz.mpg.de, magnus.nordborg@gmi.oeaw.ac.at**

Abstract

The majority of diploid organisms have polyploid ancestors. The evolutionary process of polyploidization (and subsequent re-diploidization) is poorly understood, but has frequently been conjectured to involve some form of "genome shock" — partly inspired by studies in crops, where polyploidy has been linked to major genomic changes such as genome reorganization and subgenome expression dominance. It is unclear, however, whether such dramatic changes would be characteristic of natural polyploidization, or whether they are a product of domestication. Here, we study polyploidization in *Arabidopsis suecica* (n = 13), a post-glacial allopolyploid species formed via hybridization of *A. thaliana* (n = 5) and *A. arenosa* (n = 8). We generated a chromosome-level genome assembly of *A. suecica* and complemented it with polymorphism and transcriptome data from multiple individuals of all species. Despite a divergence of ~6 Mya between the two ancestral species and appreciable differences in their genome composition, we see no evidence of a genome shock: the *A. suecica* genome is highly colinear with the ancestral genomes, there is no subgenome dominance in expression, and transposable element dynamics appear to be stable. We do, however, find strong evidence for changes suggesting gradual adaptation to polyploidy. In particular, the *A. thaliana* subgenome shows upregulation of meiosis-related genes, possibly in order to prevent aneuploidy and undesirable hmeologous exchanges that are frequently observed in experimentally generated *A. suecica*, and the *A. arenosa* subgenome shows upregulation of cyto-nuclear related processes,

possibly in response to the new cytoplasmic environment of *A. suecica,* with plastids maternally inherited from *A. thaliana*.

Introduction

Ancient polyploidization or whole-genome duplication is a hallmark of most higher-organism genomes[1,2], including our own[3,4]. While most of these organisms are now diploid and show only traces of polyploidy, there are many examples of recent polyploidization, in particular among flowering plants[5–9]. These examples are important because they allow us to study the process of polyploidization, rather than just inferring that it happened and trying to understand its evolutionary importance.

Wide-spread naturally occurring polyploid hybrids (i.e. allopolyploids), such as *Capsella bursa-pastoris* (Shepherd's Purse)[10–12], *Trifolium repens* (white clover)[13], *Brachypodium hybridum*[14,15]*, Arabidopsis kamchatica*[16], *Mimulus peregrinus*[17]*, Tragopogon miscellus* and *T. mirus*[18], demonstrate that natural polyploid species can quickly become successful, and even be deemed invasive[19]. Regardless of their eventual evolutionary success, new allopolyploid species face numerous challenges, ranging from those on a population level, such as bottlenecks[13,20] and competition with their diploid progenitors[21], to those on a genomic level, such as chromosome segregation[22–24] and changes to hybrid genome structure (e.g. chromosomal structural variants and aneuploidy[25]) and genome regulation (e.g. subgenome expression dominance[26] and the regulation of transposable elements[27]) — phenomena which may be enhanced by genomic conflicts between the newly merged subgenomes, leading to a "genome shock"[28]. In agreement with this, genomic and transcriptomic changes tied to the hybridization of two (or more) diverged genomes have been reported in resynthesized polyploids of wheat[29–35], *Brassica napus*[36–38] and cotton[39,40-37,41,42] (although resynthesized cotton appears genetically stable[43]).

The long-term importance of such rapid changes is less clear . For example, the transposable element transcription and mobilization observed in resynthesized wheat[33,44–46], is not reflected in the genome sequence of cultivated wheat[47]. However, other cultivated crop genomes, for example cotton, show instances of large structural rearrangements[5,48–50], biased gene loss[51], a spreading and proliferation of centromere repeats between subgenomes[52] and changes to the 3D genome structure[53]. Strawberry[6], peanut[8] and the mesopolyploids *B. rapa*[54] and maize[55] show evidence of subgenome dominance, while wheat[56], cotton[51] and *B. napus*[57] do not. The reasons for these differences are not understood.

An even greater source of uncertainty is whether allopolyploid crops are representative of natural polyploidization. Domestication is frequently associated with very strong "artifical" selection, which can dramatically alter the fitness landscape[58–62]. For example, large structural variants have been linked to favourable agronomic traits[63–65]. In addition, polyploid crops are generally quite recent, evolutionarily speaking.

Turning to non-domesticated species, genomic changes have been reported in natural allopolyploids like the ~80 years old *Tragopogon miscellus*[66,67], the ~140 years old *Mimulus pergrinus*[17]*,* and *Spartina anglica*[68] , which likely originated at the end of the 19th century — however, these examples are extremely recent and are more in line with the reported genomic changes in the resynthesized allopolyploids. Older natural allopolyploids, on the other hand, generally do not show signs of genomic changes after allopolyploidy. Examples of these include: white clover[13], *C. bursa-pastoris*[12,69], *A. kamchatica*[16,70]*, B. hybridium*[14] and the gymnosperm *Ephedra*[71].

Here we focus on an allopolyploid comparable in age to these examples, the highly selfing[72], *A. suecica* (2n = 4x = 26), formed through the hybridization of *A. thaliana* (2n = 10) and *A. arenosa* (2n = 2x/4x = 16/32), circa 16 kya, during the Last Glacial Maximum[20] and now widely established in northern Fennoscandia (Fig. 1a). The ancestral species diverged around 6 Mya[73], and, based on mitochondrial and chloroplast sequences, it is clear that *A. thaliana* is the maternal and *A. arenosa* the paternal parent of the hybrid[74], a scenario also supported by the fact that *A. arenosa* itself is a ploidy-variable species, so that *A. suecica* could readily be generated through the fertilization of an unreduced egg cell (2n = 2x) from *A. thaliana* by a sperm cell (n = 2x) from autotetraploid *A. arenosa*[20,75]. We have previously shown that, although *A. suecica* shows clear evidence of a genetic bottleneck[20], it shares most of its variation with the ancestral species, demonstrating that the species was formed through a hybridization and polyploidization process that involved many crosses and individuals. In order to study genomic change in *A. suecica*, we used long-read sequencing to generate a high-quality, chromosome-level assembly of a single individual, taking advantage of the fact that *A. suecica*, like *A. thaliana*, is highly selfing, making it possible to sequence naturally inbred individuals. The genome sequence was complemented by a partial assembly of a tetraploid outcrosser *A. arenosa*, and by short-read genome and transcriptome sequencing data from many individuals of all three species — including "synthetic" *A. suecica* generated *de novo* in laboratory crosses.

Results and discussion

1. The genome is conserved

We assembled a reference genome from a naturally inbred (i.e. the species is self-compatible[20,72]) *A. suecica* accession ("ASS3"), using 50x long-read PacBio sequencing (PacBio RS II). The absence of heterozygosity and the substantial (~11.6%) divergence between the subgenomes greatly facilitated the assembly. In contrast, assembling even a diploid genome of the outcrosser *A. arenosa* is complicated by high heterozygosity (nucleotide diversity around 3.5%[76]) coupled with a relatively high level of repetitive sequences (compared to the gene-rich *A. thaliana* genome). Our attempt to assemble a tetraploid *A. arenosa* individual, the result of which is also included here in addition to the genome of *A. suecica*, led to a very fragmented assembly of 3,629 contigs with an N50 of 331 Kb. In contrast, the *A. suecica* assembly has an N50 contig size of 9.02 Mb. The assembled contigs totaled 276 Mb (~90% of the 305 Mb genome size estimated by flow cytometry — see Supplementary Fig. 1; ~88% of the 312Mb genome size estimated by kmer analysis). Contigs were placed into scaffolds using high-coverage chromosome conformation capture (HiC) data and by using the reference genomes of *A. thaliana* and *A. lyrata* (here the closest substitute for *A. arenosa*) as guides. This resulted in 13 chromosome-scale scaffolds (Supplementary Fig. 2a). The placement and orientation of each contig within a scaffold was confirmed and corrected using a genetic map for *A. suecica* (see Methods, Supplementary Fig. 3, Supplementary Fig. 4). The resulting chromosome-level assembly (Fig. 1b) contains 262 Mb, and has an N50 scaffold size of 19.59 Mb. The five chromosomes of the *A. thaliana* subgenome and the eight chromosomes of the *A. arenosa* subgenome sum to 119 Mb and 143 Mb, respectively.

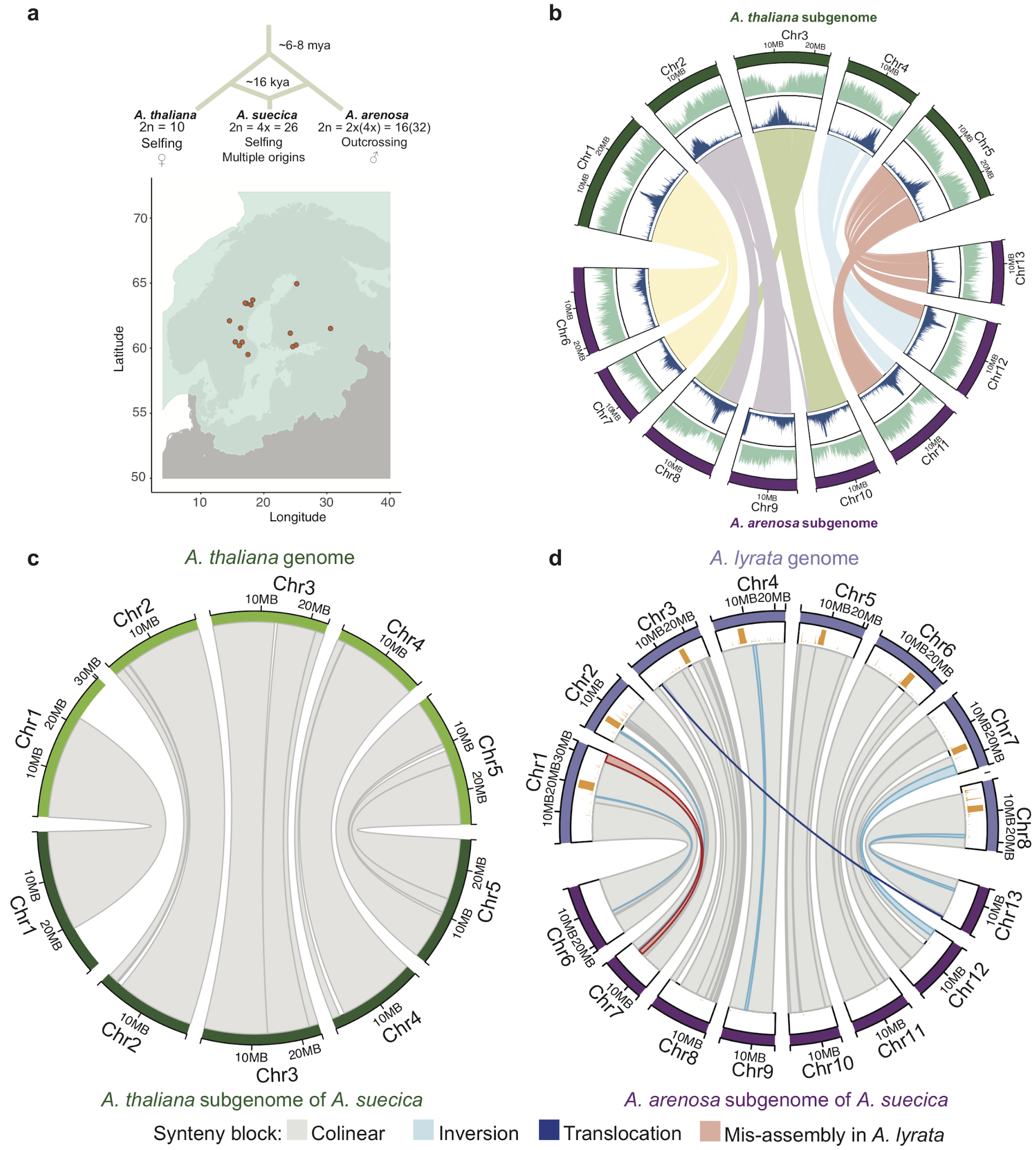


Figure 1. The genome of *A. suecica* is largely colinear with the ancestral genomes. **a** Schematic depicting the origin of *A. suecica* and its current distribution in the relation to the ice cover at the last glacial maximum (LGM). **b** The chromosome-level assembly of the *A. suecica* genome with inner links depicting syntenic blocks between the *A. thaliana* and *A. arenosa* subgenomes of *A. suecica*. The blue histogram represents the distribution of TEs along the genome and the green histogram corresponds to the distribution of protein-coding genes. **c** Synteny of the *A. thaliana* subgenome of *A. suecica* to the *A. thaliana* TAIR10 reference. In total 13 colinear synteny blocks were found. **d** Synteny of the *A. arenosa* subgenome to *A. lyrata*. In total 40 synteny blocks were found, 33 of which were colinear. Of the remaining 7 blocks, 5 represent inversions in the *A. arenosa* subgenome of *A. suecica* compared to *A. lyrata*, 1 is a translocation, and 1 corresponds to a previously reported mis-assembly in the *A. lyrata* genome[77]. Orange bars represent a density plot of missing regions ("N" bases) in the *A. lyrata* genome.

Approximately 108 and 135 Mb of the *A. thaliana* and *A. arenosa* subgenomes of *A. suecica* are in large blocks syntenic to the genomes of the ancestral species: 13 and 40 blocks, respectively (Fig. 1c,d). The vast majority of these syntenic blocks are themselves also colinear, with the exception of five small-scale inversions (~4.5 Mb) and one translocation (~244 Kb) on the *A. arenosa* subgenome— which may well (indeed probably do) reflect differences between *A. lyrata* and *A. arenosa*, two highly polymorphic species separated by about a million years[73,76]. We also corrected for the described[77] mis-assembly in the *A. lyrata* reference genome using our genetic map. Overall we find that approximately 93% of the *A. suecica* genome is syntenic to the ancestral genomes, the 13 chromosomes of *A. suecica* having remained almost completely colinear (Fig. 1c,d). This highlights the conservation of the *A. suecica* genome and contrasts with the major rearrangements that have been observed in several resynthesized polyploids[29,32,34,36] and some crops[48,50,78]. Interestingly, major rearrangements have also been observed in synthetic *A. suecica*[79], and we see clear evidence of aneuploidy in ours — a topic to which we shall return.

A total of 45,585 protein-coding genes were annotated for the *A. suecica* reference, of which 22,232 and 23,353 are located on the *A. thaliana* and *A. arenosa* subgenomes, respectively. We assessed completeness of the genome assembly and annotation with the BUSCO set for eudicots and found 2088 (98.4%) complete genes for both the *A. thaliana* and *A. arenosa* subgenomes (Supplementary Fig. 5c,d). Of the protein-coding genes, 18,023 had a one-to-one orthology between the subgenomes of *A. suecica* and 16,999 genes were conserved in a 1:1:1:1 relationship for both subgenomes of *A. suecica* and the ancestral species (using *A. lyrata* as a substitute for *A. arenosa*) (Supplementary Data 2, Supplementary Fig. 5b). We functionally annotated lineage-specific genes in *A. suecica* (i.e. genes in *A. suecica* without a reciprocal best blast hit to *A. thaliana* or *A. lyrata*) using InterPro, and only found significant enrichment in *A. thaliana* subgenome of *A. suecica* for two GO terms (GO:0008234 and GO:0015074), both of which are associated with repeat content (Supplementary Data 2). Ancestral genes not found in the *A. suecica* genome annotation were overrepresented for functional categories of plant defense response. However, checking coverage for these genes by mapping the raw *A. suecica* whole-genome resequencing data to the ancestral genomes did not confirm their loss, suggesting rather misassembly or misannotation, which is expected due to the repetitive and highly polymorphic nature of R-genes in plants.

2. The rDNA clusters are highly variable

In eukaryotic genomes, genes encoding ribosomal RNA (rRNA) occur as tandem arrays in rDNA clusters. The 45S rDNA clusters are particularly large, containing hundreds or thousands of copies, spanning millions of base pairs[80]. The nucleolus, the site of pre-ribosome assembly, forms at these clusters, but only if they are actively transcribed, and it was observed long ago that only one parent's rDNA tended to be involved in nucleolus formation in inter-specific hybrids, a phenomenon known as "nucleolar dominance"[81–8384]. In *A. suecica,* it was observed that the rDNA clusters inherited from *A. thaliana* were silenced[85,8681–8387], and structural changes associated with these clusters were also suggested[88].

Given this, we examined the composition of 45S rDNA repeats as well as their transcription. While the large and highly repetitive 45S rDNA clusters are not part of the genome assembly, it is possible to measure the copy number of *A. thaliana* and *A. arenosa* 45S rRNA genes using sequencing coverage (see Methods), and we find three accessions to have experienced massive loss of the *A. thaliana* rDNA loci (Fig. 2a), which we confirmed for one of the accessions ("AS90a") by FISH analysis (Fig. 2b,c). However, there is massive copy number variation for 45S rRNA genes in *A. suecica* (Fig. 2a), and some accessions (e.g., the reference accession "ASS3") have higher *A. thaliana* than *A. arenosa* 45S rRNA copy number (Fig. 2d,e).

Turning to expression, we also find nucleolar dominance to be variable in *A. suecica* (see Methods and Supplementary Fig. 6), with the majority of accessions expressing both 45S rRNA alleles, five exclusively expressing *A. arenosa* 45S rRNA, and one exclusively expressing *A. thaliana* 45S rRNA (Fig. 2a).

This extensive variation in 45S cluster size and expression is reminiscent of the genetically controlled intraspecific variation seen in *A. thaliana* (where different accessions express either the chromosome 2 or chromosome 4 rDNA cluster, or both[89,90]), and is in agreement with a previous observation made in natural *A. suecica* that both rDNA clusters can be expressed[91]. This suggests that the phenomenon of nucleolar dominance can at least partly be explained by retained ancestral variation. However, the large-scale decrease in rDNA cluster size observed in some accessions may be a direct consequence of allopolyploidization itself, as synthetic *A. suecica* sometimes shows immediate loss of 45S rDNA (even as early as the F1 stage) and this too varies between siblings and generations (Supplementary Fig. 6a). Elimination of rDNA loci has also been previously observed in synthetic wheat[92], and loss of rDNA sites has been reported at higher ploidy levels in strawberry[93].

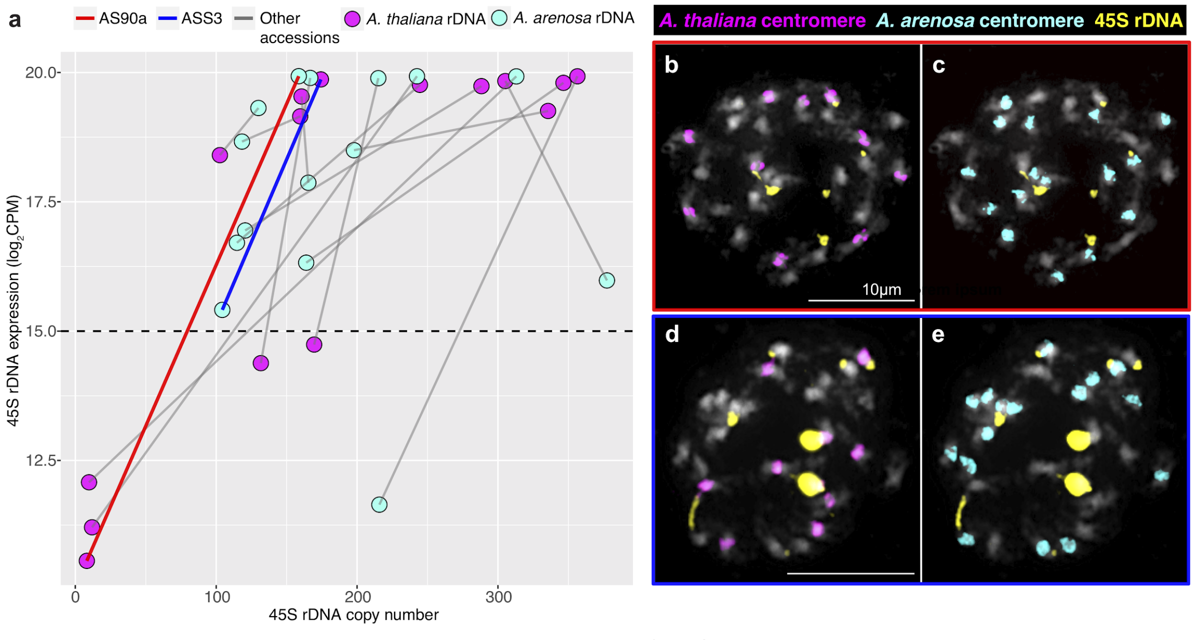


Figure 2. Expression and copy number variation of 45S rDNA in *A. suecica*. a The relationship between expression levels (log2 CPM) and copy number of 45S rDNA shows extensive variation of 45S rDNA copy number and varying direction of "nucleolar dominance". Grey lines connect subgenomes of the same accession. Values above the dashed line are taken as evidence for the expression of a particular 45S rDNA allele, as this is above the maximum level of mis-mapping seen in the ancestral species here used as a control (see Supplementary Figure 6b). **b** and **c** FISH results of a natural *A. suecica* accession AS90a that has largely lost the rDNA cluster of the *A.thaliana* subgenome (8 copies calculated for the *A. thaliana* 45S rDNA and 159 copies of the *A. arenosa* 45S rDNA). **d** and **e** FISH result of a natural accession ASS3 that has maintained both ancestral rDNA loci (174 copies calculated for the *A. thaliana* 45S rDNA and 104 copies of the *A. arenosa* 45S rDNA).

3. No evidence for abnormal transposon activity

The possibility that hybridization and polyploidization leads to a "genome shock" in the form of increased transposon activity has been much discussed[27,28,94,95]. Evidence for TE proliferation following hybridization has been found for *Ty3/gypsy* retrotransposons in hybrid sunflower species[96], though notably the hybrid sunflower species occupy habitats that are abiotically extreme[97] which is also implicated in LTR proliferation[98]. On the other hand, analysis of TE expression in F1 hybrids between *A. thaliana* and *A. lyrata* found strong correlation, even under drought stress, to the parent species, as well as little alteration of the chromatin marks H3K9me2 and H3K27me3[99] — although it remains unclear whether the F1 generation is not too early to study TE misregulation. Here we examine TE dynamics in natural *A. suecica.*

The two subgenomes of *A. suecica* differ massively in transposon content: there are almost twice as many annotated transposons in the *A. arenosa* as in the *A. thaliana* subgenome (66,722 vs 33,420; see Supplementary Figs. 5a and 7), and the true difference is almost certainly greater given that the *A. arenosa* subgenome assembly is less complete (and many of the missing regions are likely to be repeat-rich) and that the transposon annotation is biased towards *A. thaliana.* Has the combination of two such different genomes lead to increased transposon activity?

Our assembled *A. thaliana* subgenome does contain roughly 3,000 more annotated transposons than the TAIR10 *A. thaliana* reference genome, but this could reflect greater transposon number in the *A. thaliana* ancestors of this genome rather than increased transposon activity in *A. suecica.* In order to gain insight into transposon activity in *A. suecica*, we need to identify jumps that occurred after the species separated (and are thus only found in this species). We used the software PopoolationTE2[100] to call presence-absence variation on a population-scale level using genome re-sequencing datasets for 15 natural *A. suecica* accessions, 18 *A. thaliana* accessions genetically close to *A. suecica*, and 9 *A. arenosa* lines. Of the 24,569 insertion polymorphisms called with respect to the *A. thaliana* subgenome, 8,767 were shared between *A. thaliana* and *A. suecica*, 7,196 were only found in *A. thaliana*, and 8,606 were only found in *A. suecica*. Of the 115,336 insertions on the *A. arenosa* subgenome of *A. suecica,* 13,177 were shared with *A. arenosa*, 83,964 were unique to *A. arenosa*, and 18,195 were unique to *A. suecica* (Supplementary Data 1a,b; Supplementary Figs. 8,9). Considering the number of transposons per individual genome (Fig. 3a), we see that most transposon insertions in a typical *A. thaliana* subgenome are also found in *A. thaliana*, and that the slightly higher transposon load in the *A. thaliana* subgenome is mainly due to these. The reason for this is likely a population bottleneck. In contrast, the number of recent insertions (that are unique to the species) is not higher in the *A. thaliana* subgenome, suggesting that transposon activity in this subgenome is not increased.

Turning to the *A. arenosa* subgenome, we see that a typical *A. suecica* contains only about half the number of transposons of a typical *A. arenosa* individual (Fig. 3a). However, the latter is an outcrossing tetraploid, and it is thus fairer to compare with the number of transposons in four randomly chosen *A. arenosa* subgenomes of *A. suecica* (shown as "*A. arenosa* in *A. suecica* (4n)" in Fig. 3a). This largely accounts for the observed difference, but there are still clearly fewer transposons in *A. suecica*. A population bottleneck likely explains much of this, but it is impossible to rule out a contribution of decreased transposon activity in *A. suecica* as well, which might be explained by its transition to self-fertilization, which is often associated with reduced TE activity[101].

To sum up, we see no evidence for a burst of transposon activity accompanying polyploidization in *A. suecica*, a conclusion also supported by a lack of increase in transposon expression for both synthetic and natural *A. suecica* compared to the *A. thaliana* and *A. arenosa* on both subgenomes (Supplementary Fig. 9), in agreement with observations made in *A. thaliana* and *A. lyrata* F1 hybrids[99]. We do see clear traces of the population bottleneck accompanying the origin of *A. suecica*, however. The frequency distribution of polymorphic transposon insertions private in *A. suecica* is heavily skewed towards zero — almost certainly because of purifying selection because the distribution is more similar to that of non-synonymous SNPs than to that of synonymous SNPs (Fig. 3b,c). However, for both subgenomes, *A. suecica* also contains a large number of fixed or nearly-fixed insertions that are present in the ancestral species at lower frequency (Fig. 3d,e). These are likely to have reached high-frequency as a result of a bottleneck. Shared transposons are enriched in the pericentromeric regions of the genome depleted of protein-coding genes, while unique transposons insertions, which are generally at low frequency, show a more uniform distribution across the genome, consistent with evidence for stronger selection against transposon insertion in the relatively gene-dense chromosome arms[102,103] (Supplementary Fig. 10).

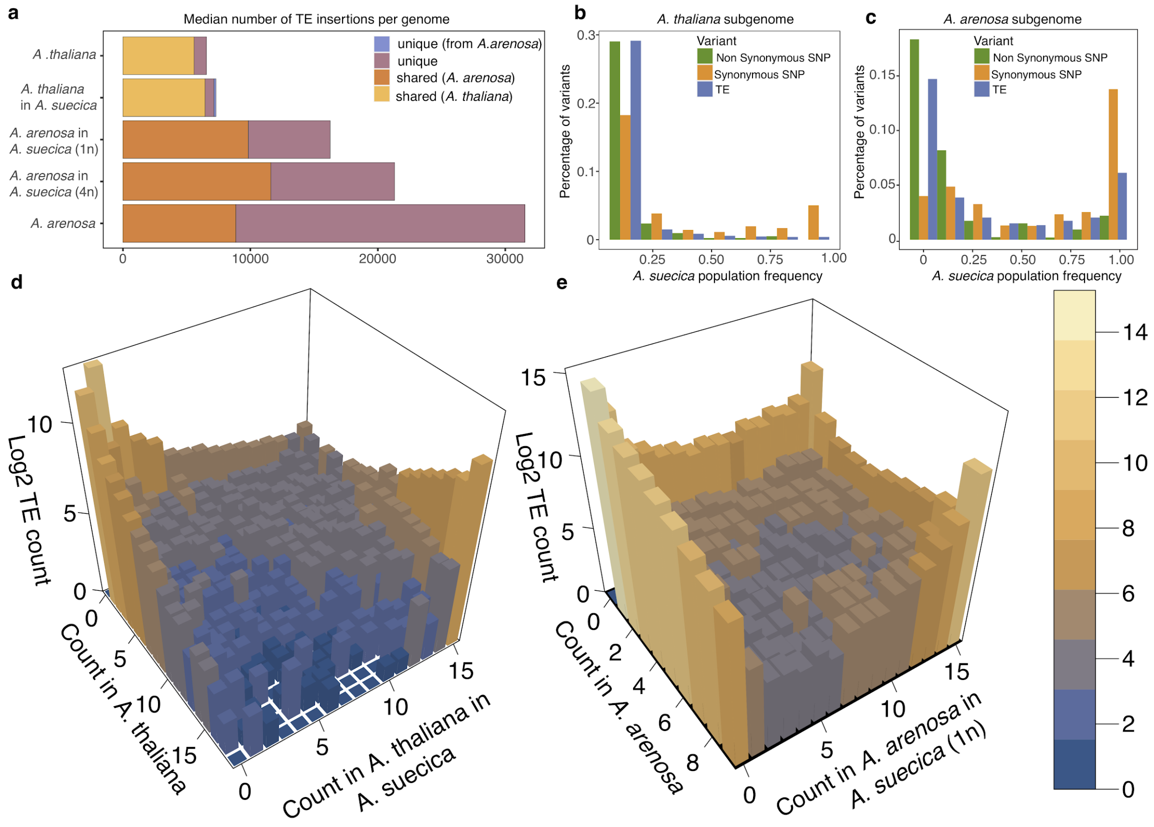


Figure 3. TE dynamics in *A. suecica* reveal no evidence for abnormal transposon activity. a Median TE insertions per genome. As the *A. arenosa* population is an autotetraploid outcrosser, 4 randomly chosen haploid *A. arenosa* subgenomes of *A. suecica* were combined to make a 4n *A. suecica*. *A. suecica* does not show an increase in private TE insertions compared with the ancestral species for both subgenomes, and shared TEs constitute a higher fraction of TEs in *A. suecica* reflecting the strong population bottleneck at its origin. Site-frequency spectra of non-synonymous SNPs, synonymous SNPs and TEs in the **b** *A. thaliana* and **c** *A. arenosa* subgenomes of *A. suecica* suggest that TEs are under purifying selection on both subgenomes. **d** 3D histogram of a joint TE frequency spectrum for *A. thaliana* on the x-axis and the *A. thaliana* subgenome of *A. suecica* on the y-axis **e** 3D histogram of a joint TE frequency spectrum for *A. arenosa* on the x-axis and the *A. arenosa* subgenome of *A. suecica* on the y-axis. **d** and **e** show stable dynamics of private TEs in *A. suecica* and a bottleneck effect on the ancestral TEs (shared) at the origin of the *A. suecica* species.

An interesting subset of recent transposon insertions unique to *A. suecica* are those that have jumped between the two subgenomes. We searched for full-length transposon copies that are present in both subgenomes of *A. suecica* and then assigned the resulting consensus sequences to either the *A. thaliana* or the *A. arenosa* ancestral genome using BLAST (see Methods). We were able to assign 15 and 56 consensus sequences as being specific to the *A. thaliana* and *A. arenosa* ancestral genome, respectively. Using these sequences, we searched our transposon polymorphism data for corresponding polymorphisms, and identified 1,515 *A. arenosa* transposon polymorphisms on the *A. thaliana* subgenome, and 496 *A. thaliana* transposon polymorphisms on the *A. arenosa* subgenome. Like other private polymorphisms, these are skewed towards rare frequencies, and are uniformly distributed across the (sub-)genome. Most of the transposons that have jumped into the *A. thaliana* subgenome are helitrons and LTR elements (Supplementary Fig. 12). LTR (copia) elements also make up most of the *A. thaliana* transposons segregating in the *A. arenosa* subgenome. The fact that roughly three times as many new insertions appear to have resulted from jumps from *A. arenosa* to *A. thaliana* than the other way around is notable. It is suggestive of higher transposon activity in the *A. arenosa* subgenome, but we have to consider differences in genome size and transposon number. If there were no differences in activity, we would expect the number of cross-subgenome jumps to be proportional to the number of potential source elements and the size of the target genome. As we have seen, the *A. arenosa* subgenome contains roughly twice as many transposons as the *A. thaliana* subgenome, but is about 20% larger. We would thus expect a 1.7-fold difference, not a three-fold one.

In conclusion, transposon activity in *A. suecica* appears to be governed largely by the same processes that governed it in the ancestral species.

4. No global dominance in expression between the subgenomes

Over time the traces of polyploidy are erased through an evolutionary process involving gene loss, often referred to as fractionation or re-diploidization[104,105106–108]. Analyses of retained homeologs in ancient allopolyploids such as *A. thaliana*[109], maize[55], *B. rapa*[54] and *Gossypium raimondii*[110] have revealed that one "dominant" subgenome remains more intact, with more highly expressed homeologs compared to the "submissive" genome(s)[109]. This pattern of "biased fractionation" has not been observed in ancient autopolyploids[111], such as pear[112], and is believed to be allopolyploid-specific.

Studying genome expression dominance in contemporary allopolyploids is useful for understanding or predicting which of the subgenomes will likely be refractory to, and which will likely experience this fractionation process more, over time[55]. Subgenome dominance in expression has been reported for a number of more recent allopolyploids such as strawberry[6], peanut[8], *Spartina*[68], *T. miscellus*[113], monkeyflower[17] and synthetic *B. napus*[114]. However, some allopolyploids display even subgenome expression, among them *C. bursa-pastoris*[10,12], white clover[13], *A. kamachatica*[70] and *B. hybridum*[14].

Subgenome dominance is often linked to differences in transposon content[6] and/or large genetic differences between subgenomes[115]. This makes *A. suecica*, with 6 Mya divergence between the gene-dense *A. thaliana* and the transposon-rich *A. arenosa*, a promising candidate to study this phenomenon at unprecedented resolution. Previous reports on subgenome dominance in *A. suecica* are conflicting, suggesting a bias to either the *A. thaliana*[116] or the *A. arenosa*[117] subgenome.

To investigate the evolution of gene expression in *A. suecica*, we generated RNA-seq data for 15 natural *A. suecica* accessions, 15 closely related *A. thaliana* accessions, 4 *A. arenosa* individuals, a synthetically generated *A. suecica* from a lab cross (the 2nd and 3rd hybrid generations) and the parental lines of this cross. Each sample had 2-3 biological replicates (Supplementary Data 2). On average, we obtained 10.6 million raw reads per replicate, of which 7.6 million reads were uniquely mapped to the *A. suecica* reference genome and 14,041 homeologous gene pairs (see Methods). On average, ~1% of A . thaliana and ~6% of *A. arenosa* RNA reads cross-mapped between the subgenomes of *A. suecica*. The ~6% of cross mapping in *A. arenosa* is likely because of the high level of polymorphism in this outcrossing species, which means that some exons in a particular *A. arenosa* individual are, in fact, closer to their homeologs on the *A. thaliana* subgenome than to those on the *A. arenosa* subgenome of *A. suecica.* However, diversity within *A . suecica* is massively lower, which means transcripts from the *A. arenosa* subgenome will almost certainly map back to the correct subgenome (see Methods and Supplementary Fig. 13).

Considering the difference in expression between homeologous genes, we found no general bias towards one or the other subgenome of *A. suecica*, for any sample or tissue, including synthetic *A. suecica* (Fig. 4a and Supplementary Fig. 14a). This strongly suggests that the expression differences between the subgenomes have not changed systematically through polyploidization, and is in contrast to previous studies, which reported a bias towards the *A. thaliana*[116] or the *A. arenosa*[117] subgenome, likely because RNA-seq reads were not mapped to an appropriate reference genome.

The set of genes that show large expression differences between the subgenomes appears not to be biased towards any particular gene ontology (GO) category, and is furthermore not consistent between accessions and individuals (Fig. 4b, Supplementary Fig. 14b,c). This suggests that many large subgenome expression differences are due to genetic polymorphisms within *A. suecica* rather than fixed differences relative to the ancestral species.

Levels of expression dominance were reported to vary across tissues in natural *C. bursa-pastoris*[11] and also resynthesized cotton[118]. To test whether expression dominance can vary for tissue-specific genes, we examined homeologous gene-pairs where at least one gene in the gene pair showed tissue specific expression, in whole-rosettes and floral buds. We do not find evidence for dominance between subgenomes in tissue specific expression either (Fig. 4b). Interestingly, the 897 genes with significant expression in whole rosettes for both homeologs showed GO overrepresentation that included both photosynthesis and chloroplast related functions (Supplementary Table 1). This result suggests that the *A. arenosa* subgenome has established important cyto-nuclear communication with the chloroplast inherited from *A. thaliana*, rather than being silenced. 2,176 gene pairs with floral bud specific expression for both homeologs were overrepresented for GO terms related to responses to chemical stimuli, such as auxin and jasmonic acid, which may reflect early developmental changes in this young tissue (Supplementary Table 1). Although flowers of selfing *A. thaliana* and *A. suecica* are scentless and are much smaller than those of the outcrosser *A. arenosa*[72], this result suggests the “selfing syndrome”[119] has not hugely impacted the transcriptome of floral buds in *A. suecica*, at least at this stage of development.

In summary, we find no evidence that one subgenome is dominant and contributes more to the functioning of *A. suecica*. On the contrary, homeologous gene pairs are strongly correlated in expression across tissues.

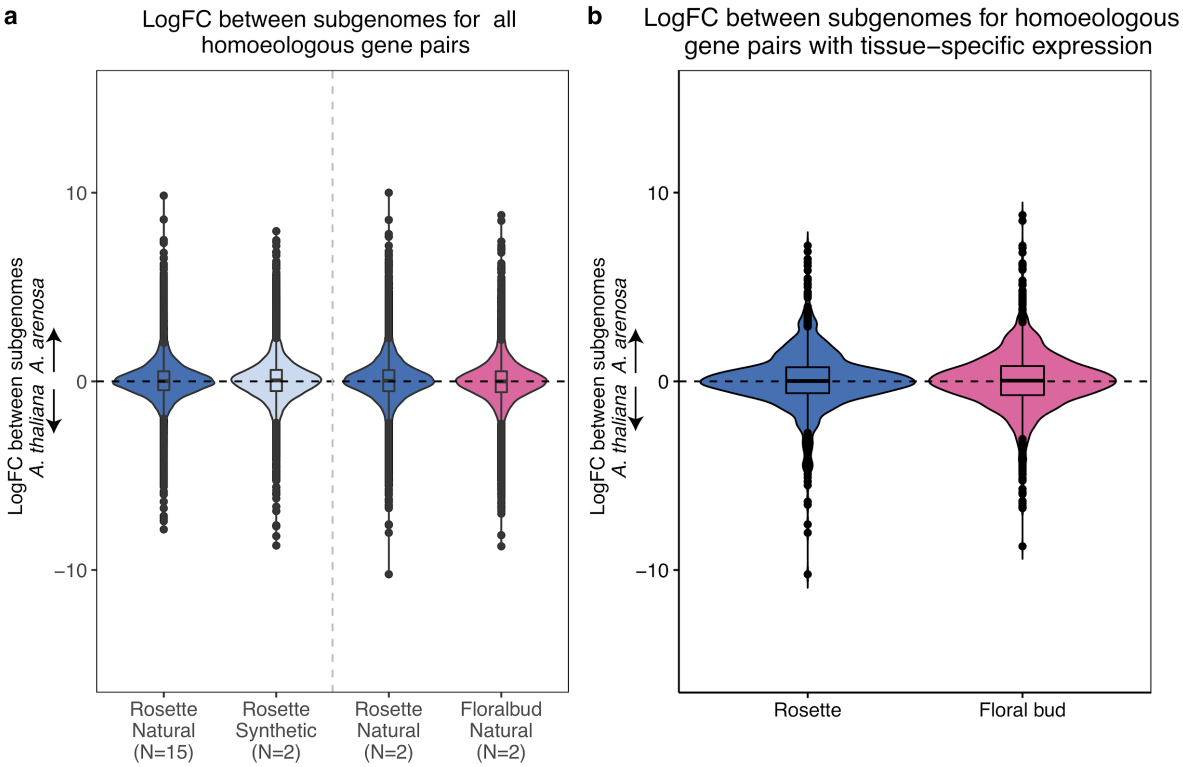


Figure 4. Patterns of gene expression between the subgenomes of *A. suecica* in rosettes and floral buds. **a** Violin plots of the mean log fold-change between the subgenomes for the 15 natural *A. suecica* accessions and two synthetic lines for whole rosettes. Mean log fold-change for the two accessions (“ASS3” and “AS530”) where transcriptome data for both whole rosettes and flower buds were available. All the distributions are centered around zero suggesting even subgenome expression. **b** Violin plots for the mean log fold-change between the subgenomes for genes with tissue-specific expression. At least one gene in a homeologous gene pair was required to show tissue-specific expression.

5. Evolving gene expression in *A. suecica*

The previous section focused on differences in expression between the subgenomes, between homeologous copies of the same gene within the same individual. This section will focus on differences between individuals, between homologous copies of genes that are part of the same (sub-)genome. To provide an overview of expression differences between individuals we performed a principal component analysis (PCA) on gene expression separately for each (sub-)genome. For both subgenomes, the first principal component separates *A. suecica* from the ancestral species and the synthetic hybrid (Fig. 5a,b, Supplementary Fig. 15), suggesting that hybridization does not automatically result in large-scale transcriptional changes, and that altered gene expression changes in natural *A. suecica* have evolved over time. Given the limited time involved, and the fact the genes that have changed expression are far from random with respect to function (Fig. 5c), we suggest that the first principal component primarily captures trans-regulated expression changes in *A. suecica* that are likely adaptive.

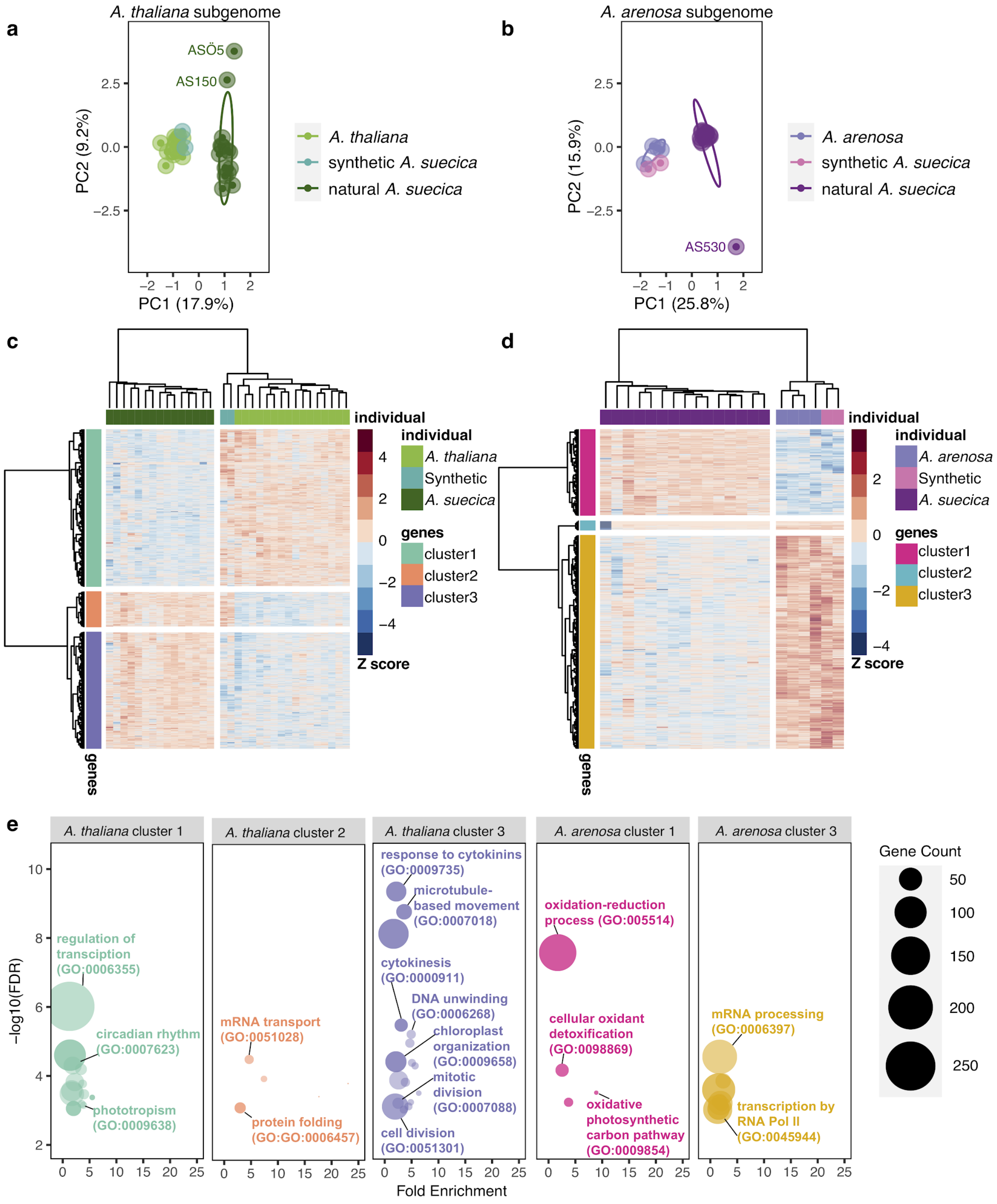


Figure 5. Differential gene expression analysis in *A. suecica*. Patterns of differential gene expression in *A. suecica* support adaptation to the whole-genome duplication for the *A. thaliana* subgenome and adaptation to the new plastid environment for the *A. arenosa* subgenome. **a** PCA for *A. thaliana*, *A. thaliana* subgenome of natural and synthetic *A. suecica* lines. PC1 separates natural *A. suecica* from the ancestral species and the synthetic lines. **b** PCA for *A. arenosa*, *A. arenosa* subgenome of natural and synthetic *A. suecica* lines. PC1 separates natural *A. suecica* from the ancestral species and the synthetic lines, whereas PC2 identifies outlier accessions discussed further below (see Fig. 6). **c, d** Heatmap of differentially expressed genes (DEGs) for the two subgenomes of *A. suecica*. Positive numbers (red color) indicate higher expression. Genes and individuals have been clustered based on similarity in expression, resulting in clusters discussed in the text. **e** Gene ontology enrichment for each cluster in **c** and **d**. Categories discussed in the text are highlighted.

To further characterize expression changes in natural *A. suecica* we analyzed differentially expressed genes (DEGs) on both subgenomes compared to the corresponding ancestral species. The total number of DEGs was 4,186 and 4,571 genes for the *A. thaliana* and *A. arenosa* subgenomes, respectively (see Methods, Supplementary Data 2). These genes were clustered based on the pattern of change across individuals (Fig. 5c,d) and GO enrichment analysis was carried out for each cluster (Fig. 5e, Supplementary Table 2).

For the *A. thaliana* subgenome, we identified three clusters. Cluster 1 comprised 2,135 genes that showed decreased expression in *A. suecica* compared to *A. thaliana*. These genes are strongly enriched for transcriptional regulation, which may be expected as we are examining DEGs between the species. Also notable are enrichments for circadian rhythm function and phototropism, which may be related to the ecology of *A. suecica* and its post-glacial migration to the Fennoscandinavia region (Fig. 1a).

Cluster 2 consisted of 468 genes that are over-expressed in both natural and synthetic *A. suecica* relative to *A. thaliana*. These expression changes are thus most likely an immediate consequence of hybridization presumably reflecting trans-regulation. Genes in this cluster are enriched for "mRNA transport" and "protein folding". The importance of the adjustment of protein homeostasis has been reported previously in experimentally evolved stable polyploid yeast[120]. Notably, the synthetic lines used in the expression analysis were selected to be healthy-looking, and did not show signs of aneuploidy (Supplementary Fig. 17).

Cluster 3 consisted of 1,583 genes that show increased expression in *A. suecica* compared to *A. thaliana*, and several of the enriched GO categories, such as microtubule-based movement, cytokinesis, meiosis and cell division, suggest that the *A. thaliana* subgenome of *A. suecica* is adapting to polyploidy at the level of basic cell biology. That there has been strong selection for this seems likely given that aneuploidy is frequent in synthetic *A. suecica* (Supplementary Fig. 16), while natural *A. suecica* has a stable and conserved karyotype. Importantly, there is independent evidence for adaptation to polyploidy via modifications of the meiotic machinery in the other ancestor of *A. suecica*, *A. arenosa*, as well[23,121,122], although we see very little overlap in the genes involved (Supplementary Fig. 16). The nature of these changes in the *A. thaliana* subgenome of *A. suecica* will require further investigation, but we note that there is enrichment (see Methods, Supplementary Data 2) for Myb family transcription factor binding sites[123] among upregulated genes in cluster 3.

For the *A. arenosa* subgenome, we also found three clusters of DEGs (Fig. 5d) with GO enrichment for two of them (Fig. 5e, Supplementary Table 2). Cluster 1 consisted of 1,278 genes that show increased expression in natural *A. suecica* compared to *A. arenosa* and synthetic *A. suecica,* and are enriched for plastid-related functions, including oxidation-reduction and the oxidative photosynthetic carbon pathway. We hypothesize that this may be due to selection on the *A. arenosa* subgenome to restore communication with the new plastid environment as plastid genomes were maternally inherited from *A. thaliana*. We also examined genes that show structural evidence for direct plastid-nuclear interactions in *A. thaliana* using CyMIRA[124]. Out of a total of 69 genes, 12 overlap genes identified in Cluster1, more than expected by chance (p-value 0.0072; one-sided Fisher Exact Test, one sided; Supplementary Data 2). Cluster 3 consists of 3,166 genes that show decreased gene expression in *A. suecica* compared to *A. arenosa* and synthetic *A. suecica*. These genes were primarily enriched for mRNA processing and epigenetic regulation of gene expression (Supplementary Table 2) and positive regulation of transcription by RNA polymerase II, which might suggests differences in the epigenetic regulation of expression between *A. arenosa* and

A. suecica. Cluster 2 (127 genes), finally, did not have a GO overrepresentation and showed an intriguing pattern discussed in the next section.

6. Homeologous exchange contributes to variation in gene expression

The second principal component for gene expression identified three outlier-accessions of *A. suecica*, two for the *A. thaliana* subgenome (Fig. 5a) and one for the *A. arenosa* subgenome (Fig. 5b). While closely examining the latter accession, "AS530", we realized that it is responsible for the cluster of genes with distinct expression patterns but no GO enrichment just mentioned (Fig. 5d, Cluster 2). Genes from this cluster were significantly downregulated on the *A. arenosa* subgenome (Fig. 6a) and upregulated on the *A. thaliana* subgenome (Fig. 6b) — for AS530 only. The further observation that 104 of the 127 genes (Supplementary Fig. 20a) in the cluster are located in close proximity in the genome, pointed to a structural rearrangement. The lack of DNA sequencing coverage on the *A. arenosa* subgenome around these 104 genes and the doubled coverage for their homeologs on the *A. thaliana* subgenome, strongly suggested a homeologous exchange (HE) event resulting in AS530 carrying four copies of the *A. thaliana* subgenome and zero copies of the *A. arenosa* genome with respect to this this, roughly 2.5 Mb region of the genome (Fig. 6c). This explanation was further supported by HiC data, which showed clear evidence for interchromosomal contacts between *A. thaliana* subgenome chromosome 1 and *A. arenosa* subgenome chromosome 6 around the breakpoints of the putative HE in AS530 (Fig. 6 d,e), and by multiple discordant Illumina paired-end reads at the breakpoints between the homeologous chromosomes, which independently support the HE event (Supplementary Fig. 19a-d).

Based on this we examined the two outlier *A. suecica* accessions for the *A. thaliana* subgenome (Fig. 5a; "AS150" and "ASÖ5"), and found that they likely share a single HE event in the opposite direction (four copies of the *A. arenosa* subgenome and no copies of the *A. thaliana* subgenome for a region of roughly 1.2Mb in size, see Supplementary Figure 18). This demonstrates that HE occurs in *A. suecica* and contributes to the intraspecific variation we observed in gene expression (Fig 5a, b). HE in allopolyploids is a main source of diversity, causing phenotypic changes in flower color in synthetic polyploid peanut[9] and extensive phenotypic change in synthetic polyploid rice at a population level[125]. However, the majority of HEs are probably deleterious as they will lead to gene loss: although the *A. thaliana* and *A. arenosa* genomes are largely syntenic, AS530 is missing 108 genes (Supplementary Figure 19) that are only present on the *A. arenosa* subgenome segment that has been replaced by the homeologous segment from the *A. thaliana* subgenome, and AS150/ASÖ5 are missing 53 genes that were only present on the *A. thaliana* subgenome.

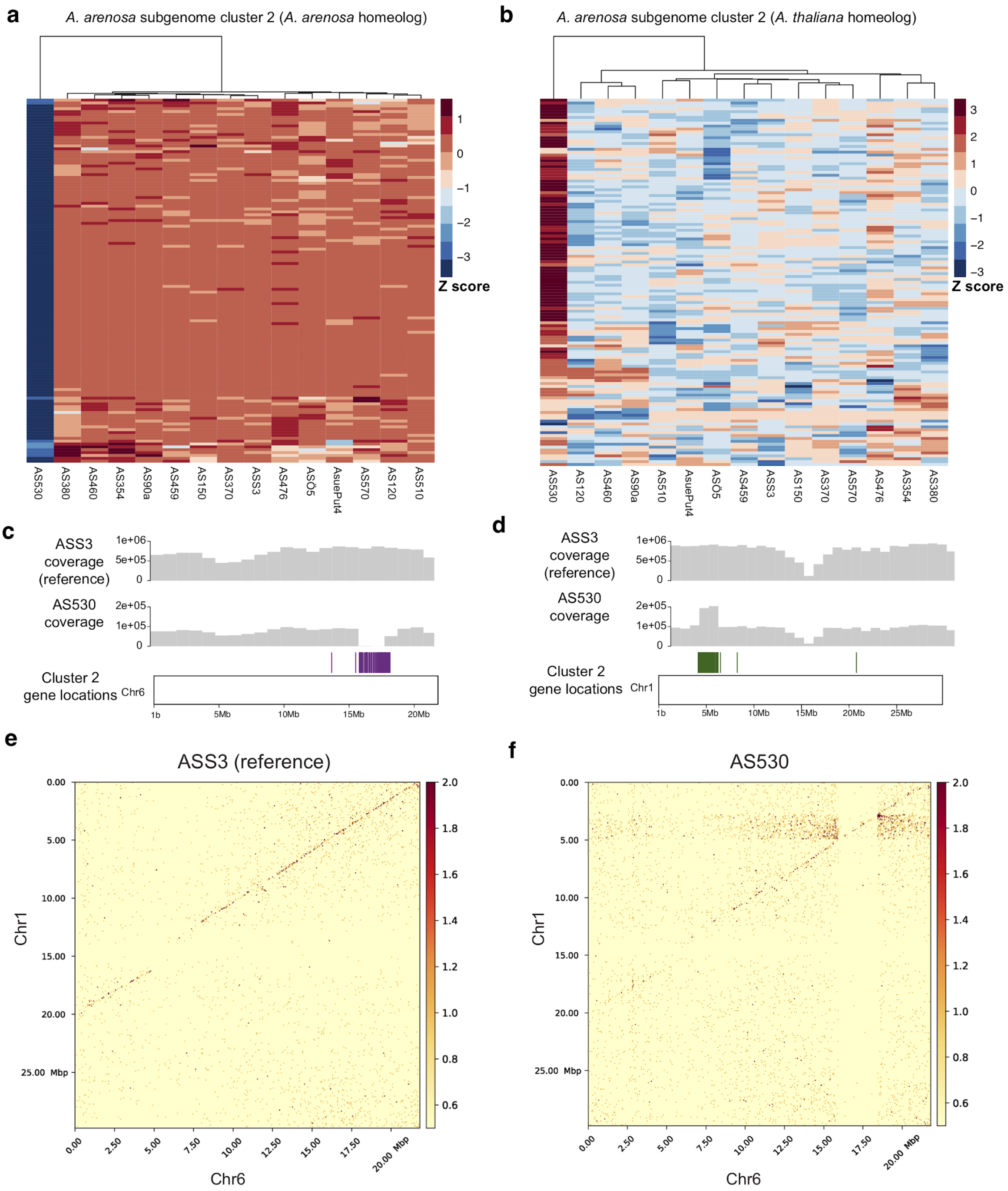


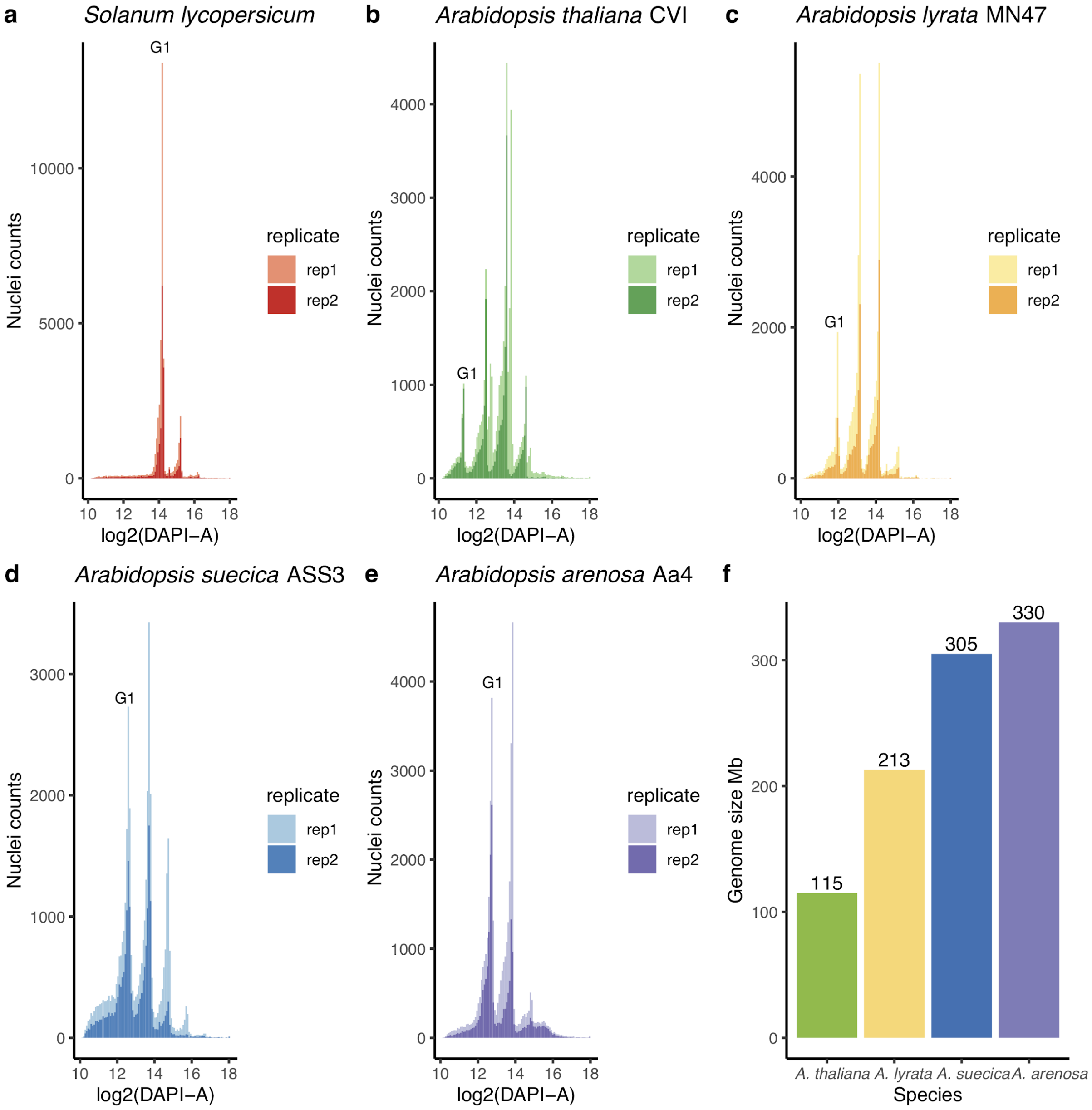
Figure 6. Homeologous exchange contributes to expression variance within *A. suecica*. a Cluster 2 of Fig. 5d explains the outlier accession AS530 which is not expressing a cluster of genes on the *A. arenosa* subgenome. **b** Homeologous genes of this cluster on the *A. thaliana* subgenome of *A. suecica* show the opposite pattern and are more highly expressed in AS530 compared to the rest of the population. **c** 97 of the 122 genes from cluster 3 are located in close proximity to each other on the reference genome but appear to be deleted in AS530 based on sequencing coverage. **d** The *A. thaliana* subgenome homeologs have twice the DNA coverage, suggesting they are duplicated. **e** HiC data show (spurious) interchromosomal contacts at 25 Kb resolution between chromosome 1 and chromosome 6 around the breakpoint of the cluster of 97 genes in AS530 but not in reference accession ASS3.

Conclusion

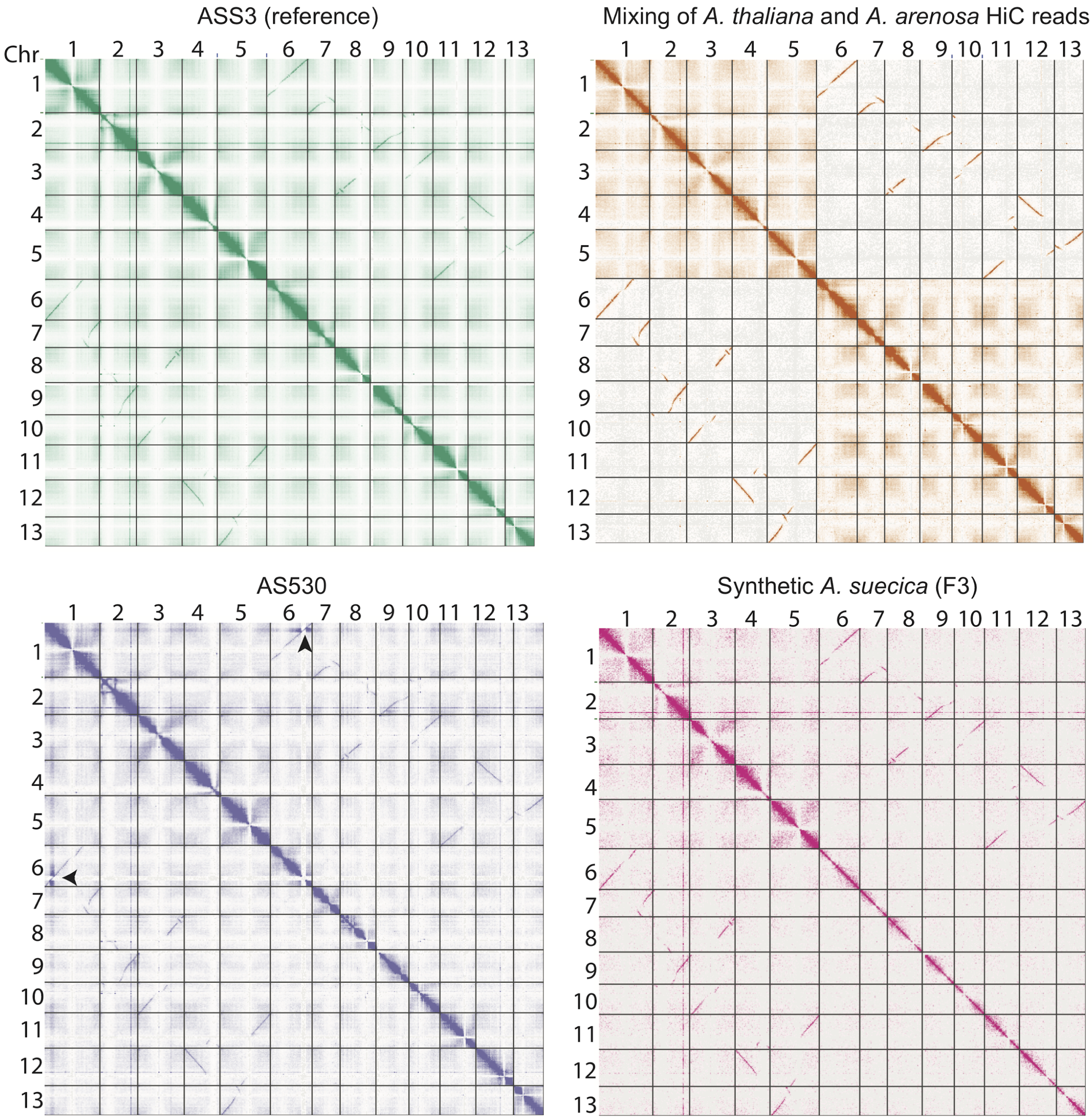
This study has focused on the process of polyploidization in a natural allotetraploid species, *A. suecica*, generated roughly 16 kya through the hybridization of two species, *A. thaliana* and *A. arenosa*, which differ substantially in everything from genome size and chromosome number to mating system and ecology. Our study is one of a growing number of studies focusing on natural rather than domesticated polyploid, but is unparalleled in its resolution thanks to one of the parents being a major model species.

Our main conclusion from this study is that polyploid speciation, at least in this case, appears to have been a gradual process rather than some kind of "event". We confirmed previous results that genetic polymorphism is largely shared with the ancestral species, demonstrating that *A. suecica* did not originate through a single unique hybridization event, but rather through multiple crosses[20]. We also find no evidence for "genome shock" (i.e. major genomic changes linked to structural and functional changes) that has often been suggested to accompany polyploidization and hybridization. The genome has not been massively rearranged, transposable elements are not out of control, and there is no subgenome dominance in expression. On the contrary, we find evidence of genetic adaptation to "stable" life as a polyploid, in particular changes to the meiotic machinery and in interactions with the plastids. These findings made in natural *A. suecica*, together with the observation that experimentally generated *A. suecica* are often unviable and do exhibit evidence of genome rearrangements, similar to the young allopolyploid species in *Tragapogon* and monkeyflower, suggest that the most important bottleneck in polyploid speciation may be selective. If this is true, domesticated polyploids may not always be representative of natural polyploidization, because of human intervention. Darwin famously argued that "Natura non facit saltum"[126] — we suggest that natural polyploids are no exception from this, but note that many more species will have to be studied before it is possible to draw general conclusions.

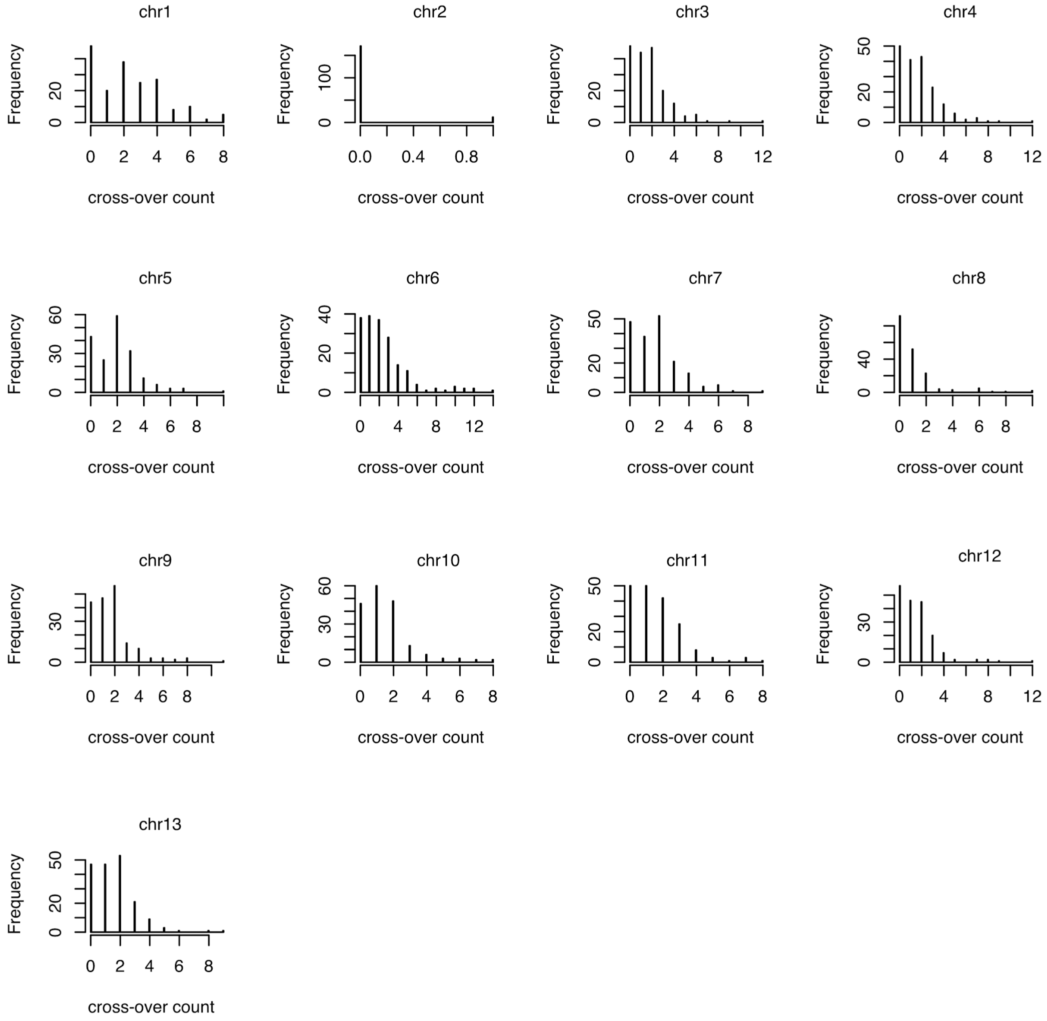
Supplemental figures



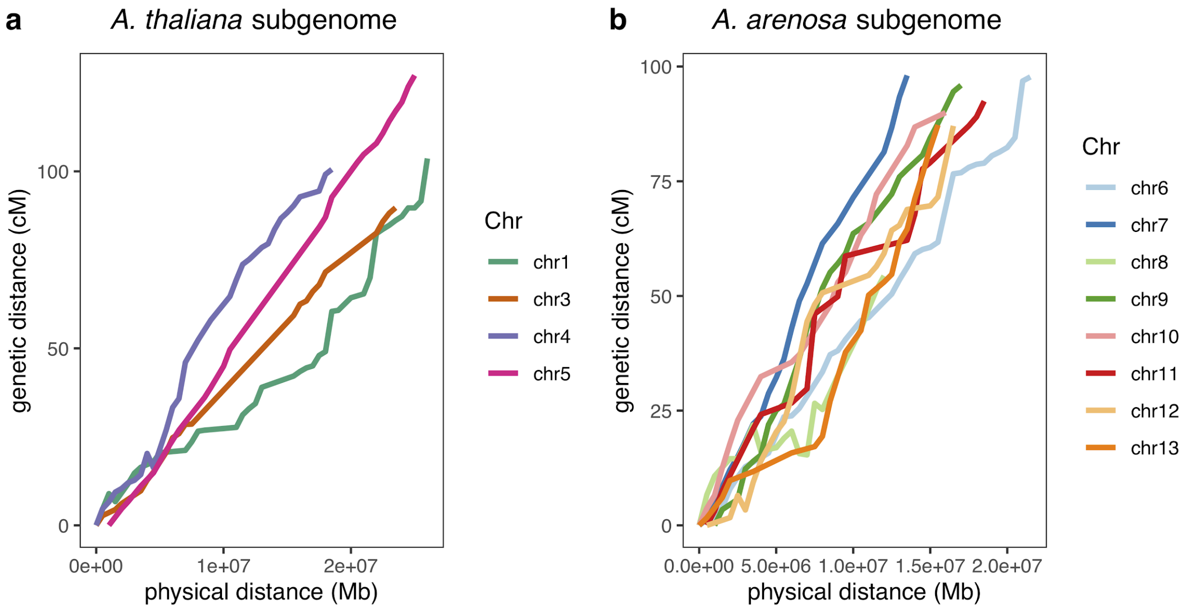
Supplementary Figure 1. Measuring genome sizes of *Arabidopsis* species using flow cytometry. a FACs sorting of *Solanum lycopersicum* cells from 3 week old leaf tissue for two replicates. G1 represents the peak denoting the G1 phase of the cell cycle. Cells in the G1 phase have 2C DNA content (i.e. a 2N genome). **b** *A. thaliana* “CVI” accession **c** *A. lyrata* “MN47” (the reference accession) **d** *A. suecica* “ASS3” (the reference accession) **e** autopolyploid *A. arenosa* accession “Aa4” **f** Bar chart shows calculated genome sizes (rounded to the nearest whole number) for each species using *Solanum lycopersicum* as the standard .



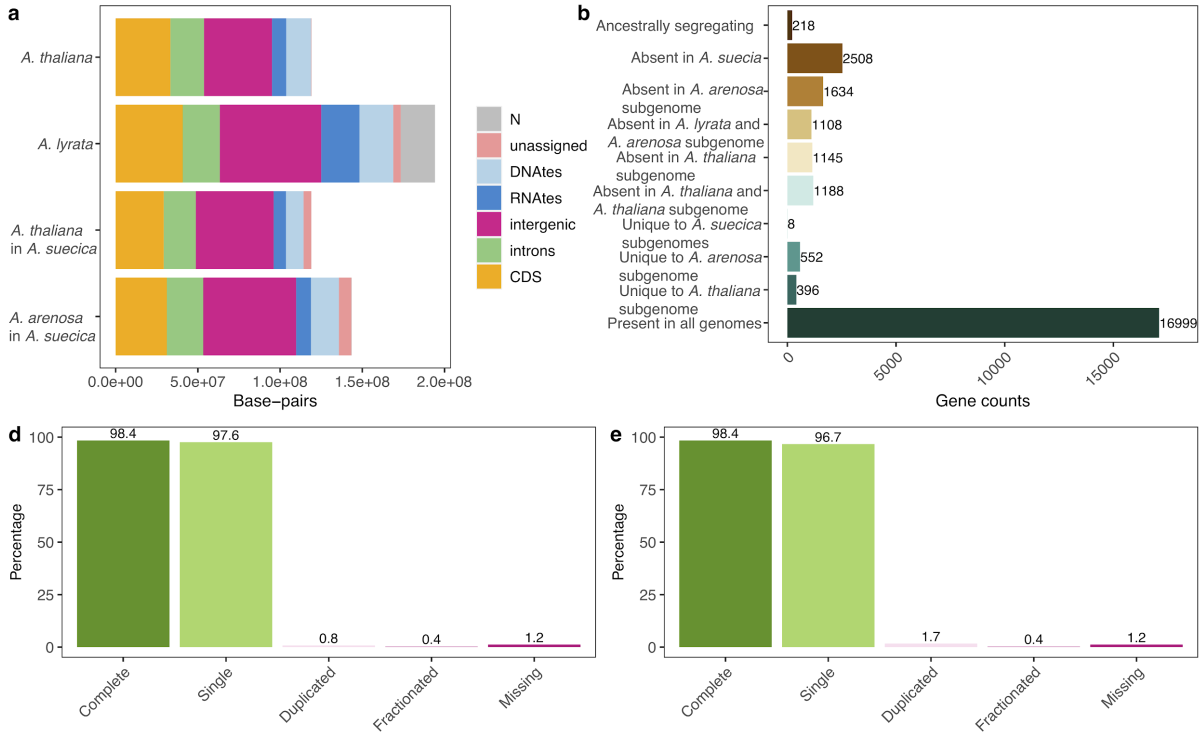
Supplementary Figure 2. HiC as a tool to investigate structural rearrangements. **a** HiC contact map for the full chromosome-level genome assembly of *A. suecica*. **b** Mixing of *A. thaliana* and *A. arenosa* HiC reads suggest interchromosomal contacts between homeologous chromosomes is a result of mis-mapping for HiC reads. Such mis-mapping is typically filtered out in short read DNA and RNA datasets using insert size and proper pairs mapping filters, however in HiC long range chromosomal contacts are not filtered out. **c** Accession “AS530” with the region of homeologous exchange highlighted with an arrow (Figure 6), no other rearrangements were observed. **d** HiC of synthetic *A. suecica* (F3).



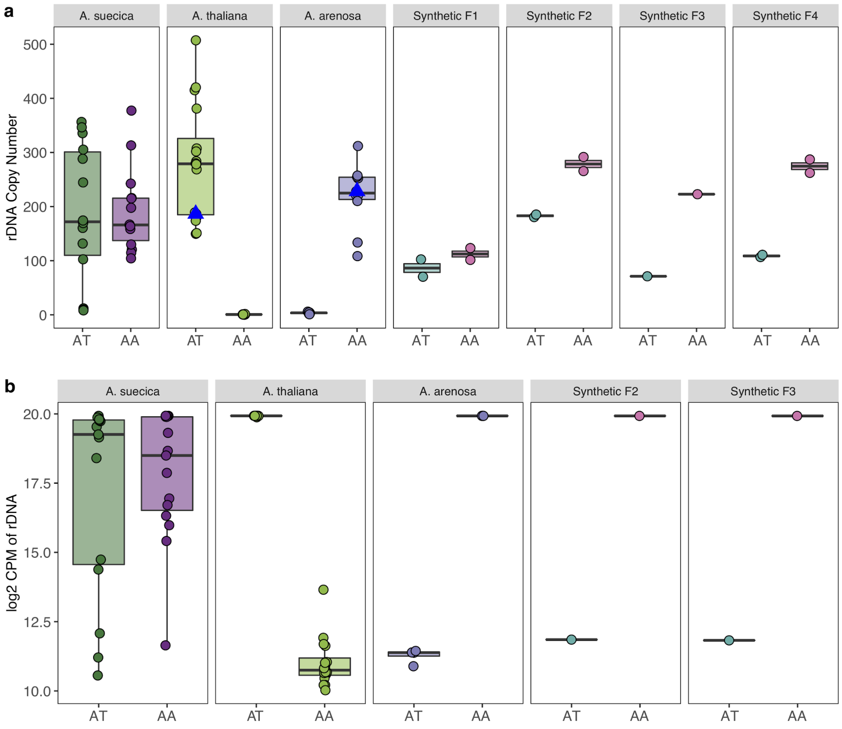
Supplementary Figure 3. Crossover counts in an *A. suecica* F2 population. Per chromosome crossover counts in our F2 population (N=185). Chromosome 2 had too few SNPs to be analysed in our cross due to the recent bottleneck in *A. suecica*[20].



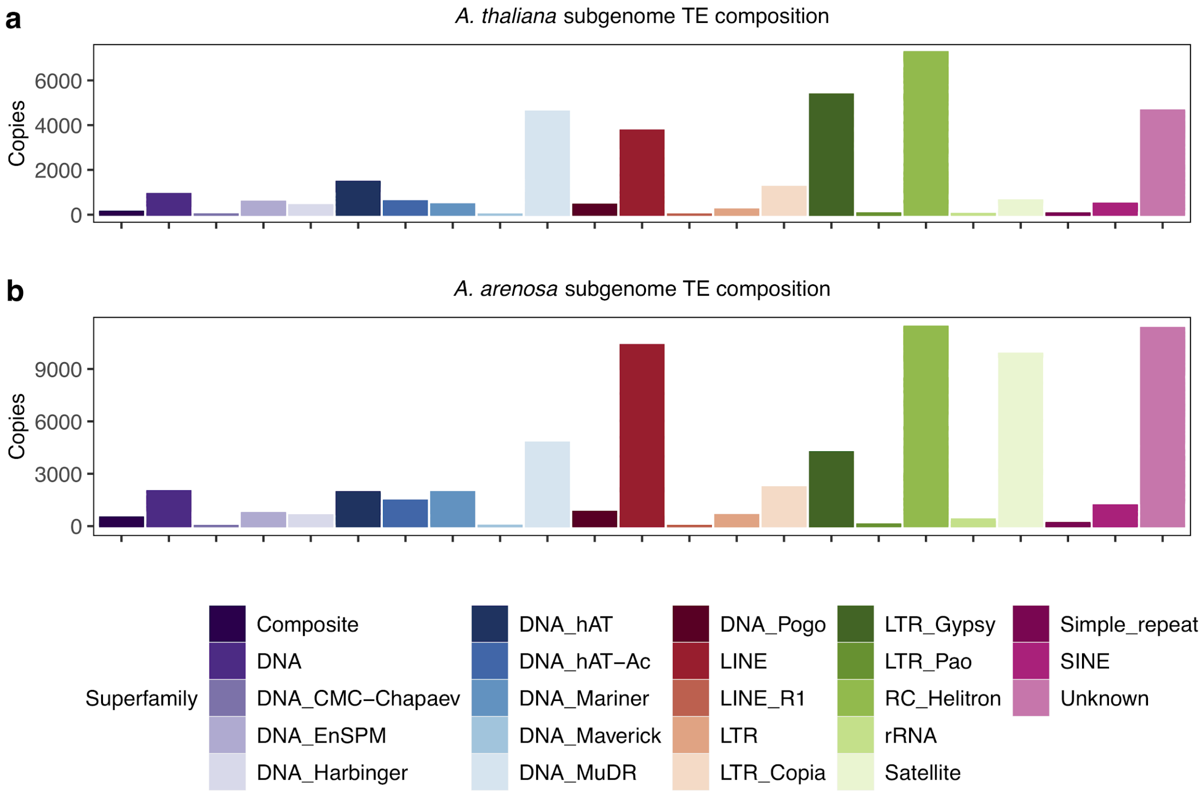
Supplementary Figure 4. A genetic map for *A. suecica*. Physical distance (Mb) vs genetic distance (cM) is plotted for each: **a** *A. thaliana* subgenome and; **b** *A. arenosa* subgenome chromosome. Chromosome 2 is not plotted as there are too few SNPs on this chromosome in our cross, due to the recent bottleneck in *A. suecica*[20]



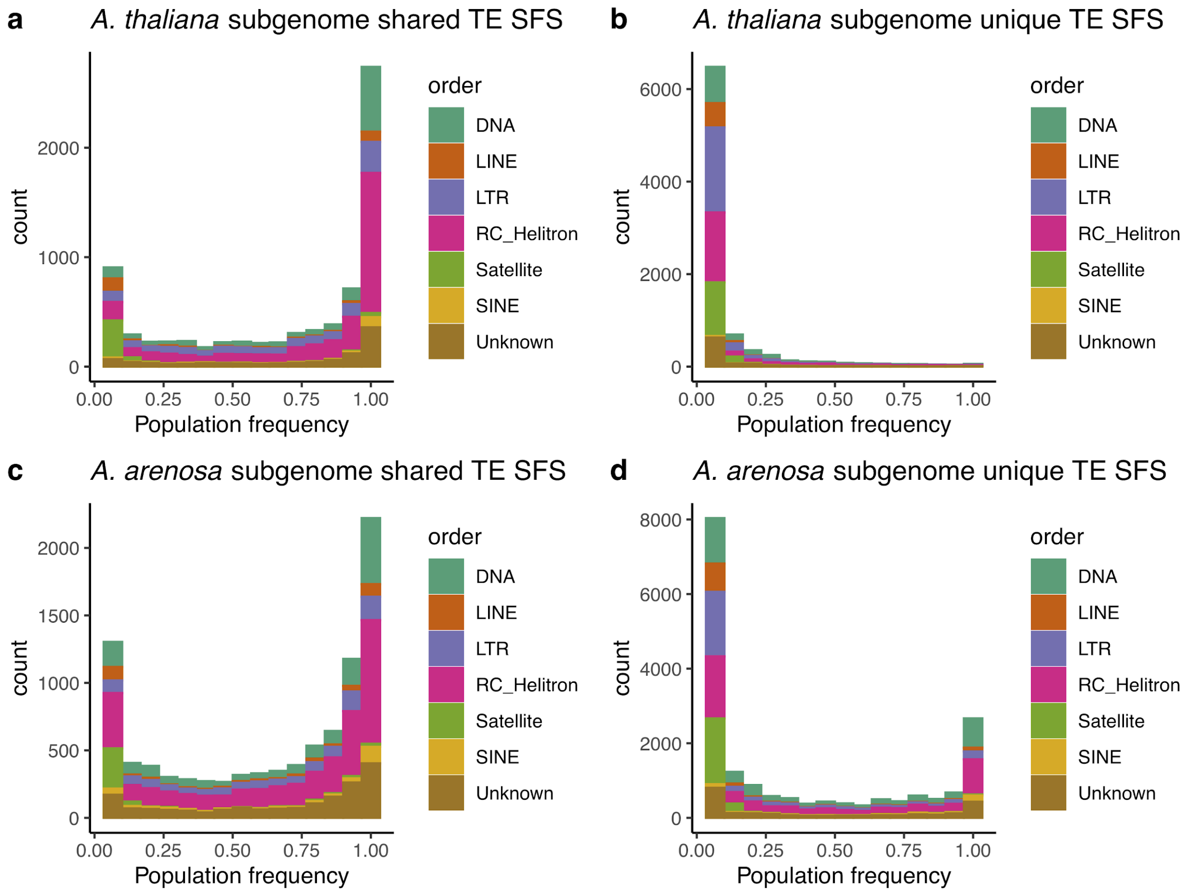
Supplementary Figure 5. Genome composition and orthologous gene relationships in *A. suecica*. a Genome composition of the *A. suecica* subgenomes and the ancestral genomes of *A. thaliana* and *A.lyrata* (here a substitute reference for *A. arenosa* because it is annotated). **b** Counts of orthologous relationships between the subgenomes of the reference *A. suecica* genome and the reference *A. thaliana* and *A. lyrata* genome. Ancestrally segregating genes are genes that are shared between the *A. thaliana* reference and the *A. arenosa* subgenome or shared between the *A. lyrata* reference and the *A. thaliana* subgenome. Therefore they most likely represent genes ancestrally segregating in the ancestor of *A. thaliana* and *A. lyrata*. BUSCO analysis of *A. suecica* using the BUSCO set for eudicots for the **d** *A. thaliana* and **e** *A. arenosa* subgenome.



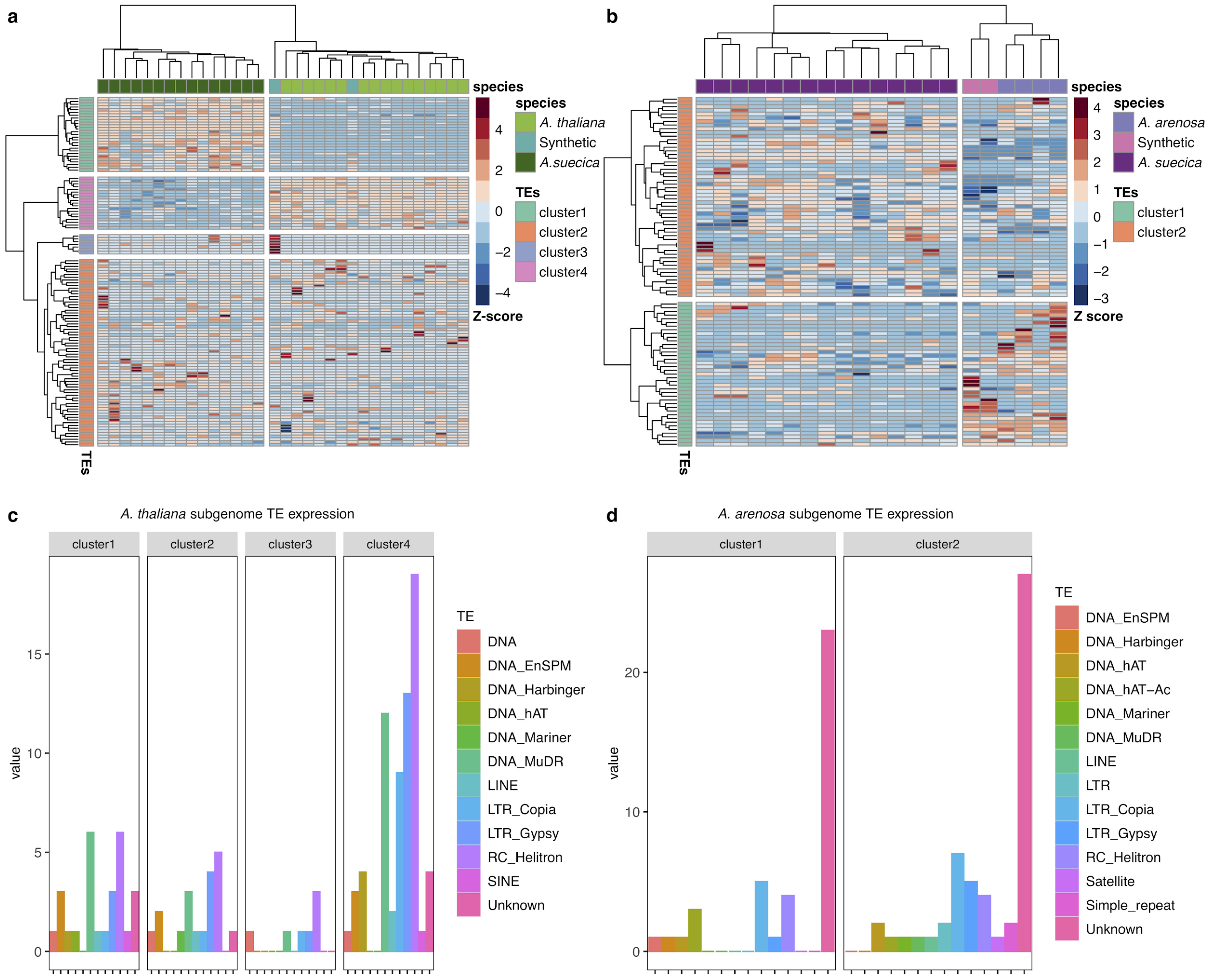
Supplementary figure 6. **rDNA copy number variation and expression. a** Copy number of *A. thaliana* and *A. arenosa* rDNA in natural *A. suecica*, ancestral species and synthetic lines. Blue triangles represent the *A. thaliana* and *A.arenosa* parent lines of the synthetic *A. suecica* cross. AT represents results when mapping to the *A. thaliana* consensus sequence and AA to the *A. arenosa* consensus sequences for the 45S rRNA **b** Expression (log2 CPM) of *A. thaliana* and *A. arenosa* rDNA in natural *A. suecica*, ancestral species and synthetic lines. Accessions with log2 CPM of >=15 was taken as evidence for expression for the *A. thaliana* and *A. arenosa* 45S rRNA in *A. suecica*, as this CPM value was above the maximum level of mis-mapping observed in the ancestral species (*A. thaliana* mapping to the *A. arenosa* 45S rRNA).



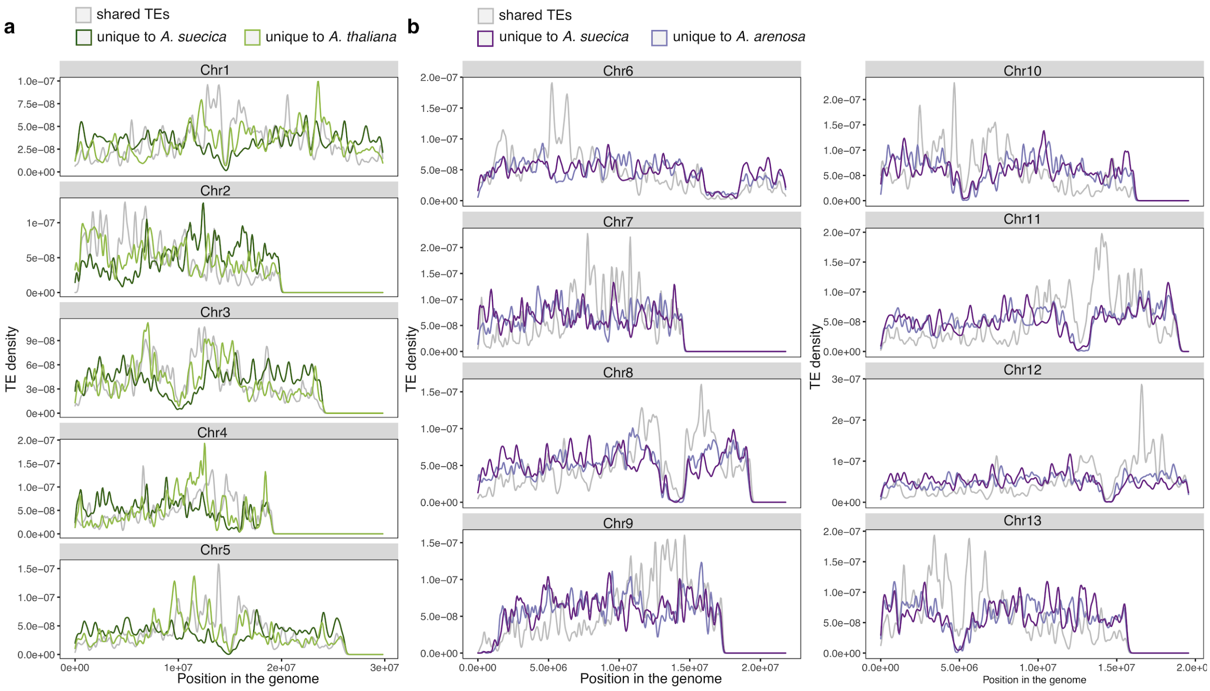
Supplementary Figure 7. TE-composition of the *A. suecica* reference genome. TE composition of the **a** *A. thaliana* and **b** *A. arenosa* subgenome of *A. suecica*.



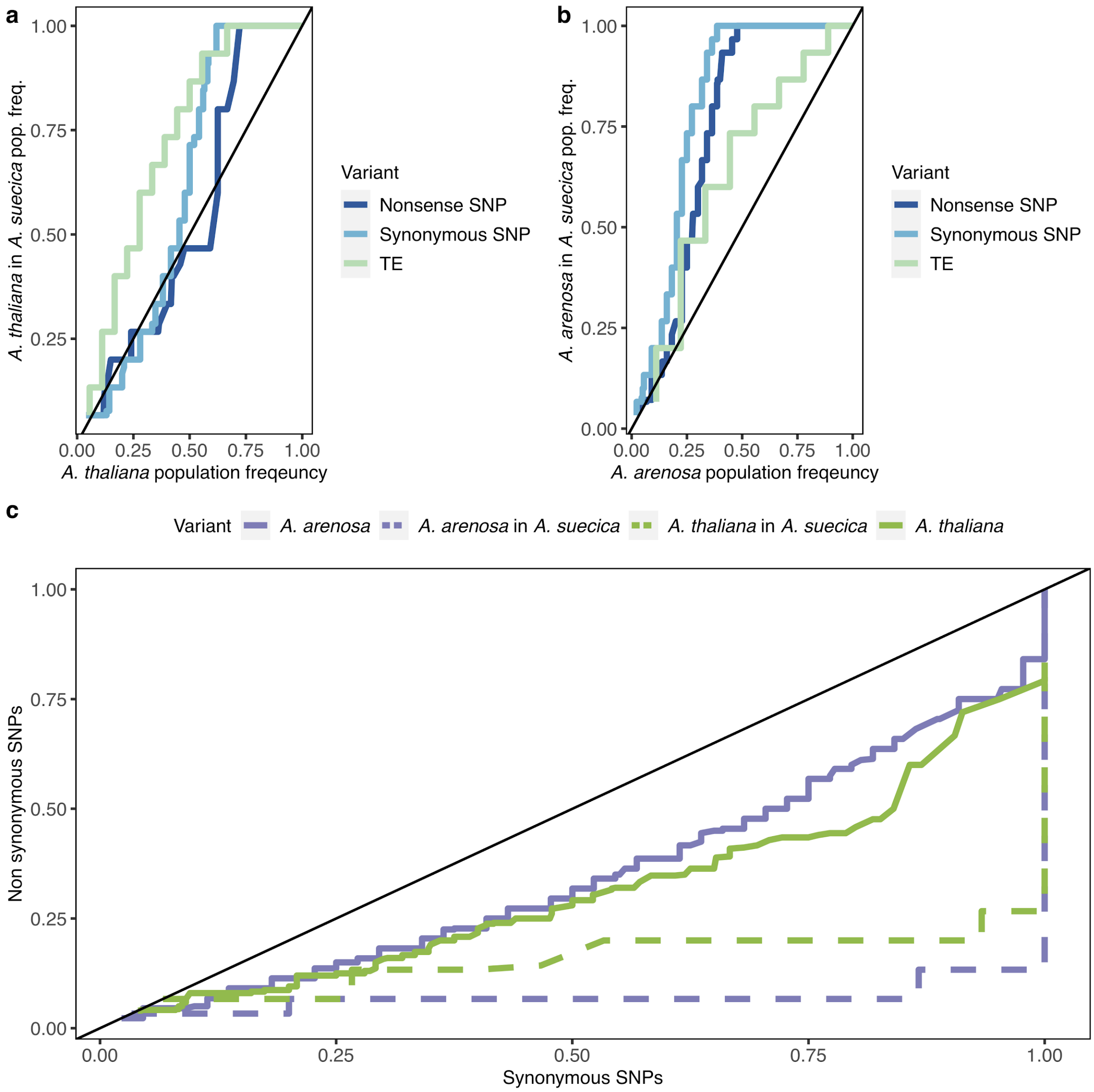
Supplementary Figure 8. Site frequency spectrum (SFS) of shared TEs and unique TEs in *A. suecica* broken down by TE family. Shared TE SFS for the **a** *A. thaliana* and **b** *A. arenosa* subgenome. Private TE SFS for the **c** *A. thaliana* and **d** *A. arenosa* subgenome.



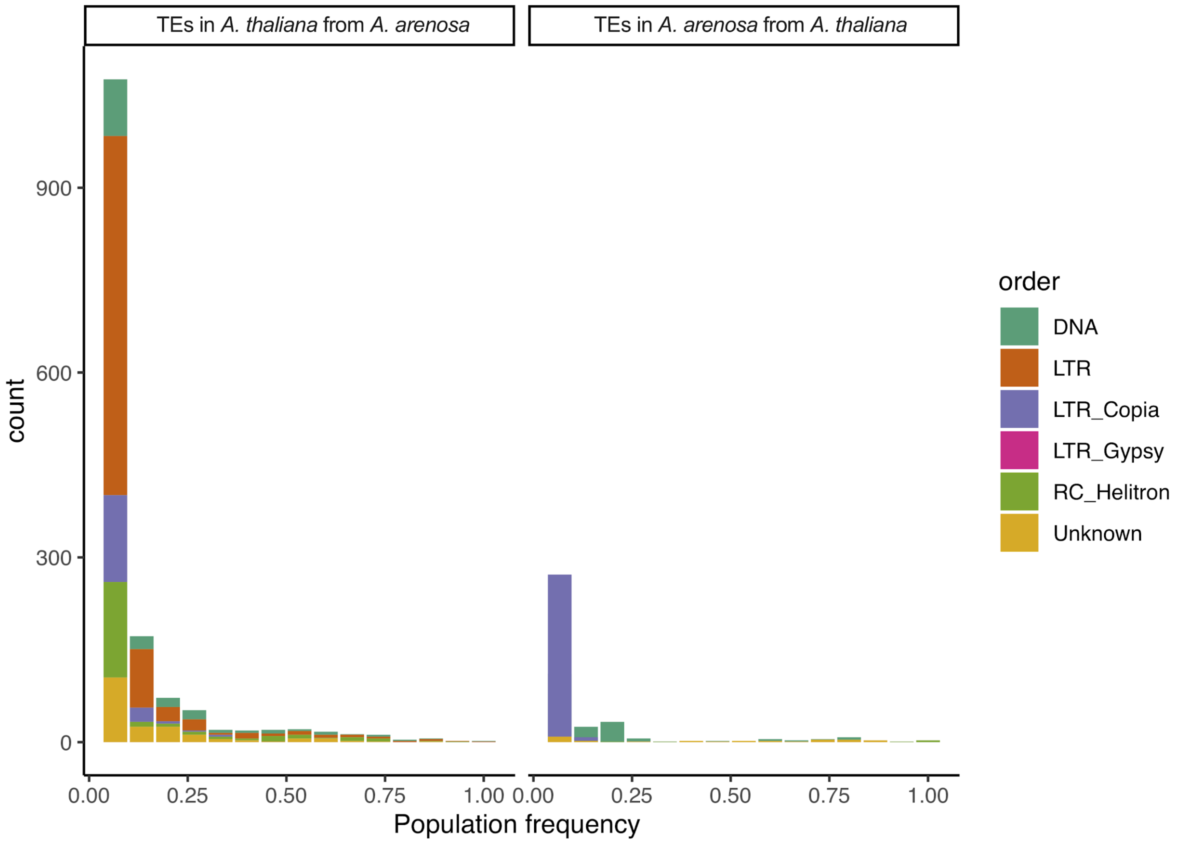
Supplementary Figure 9. Analysis of TE expression in *A. suecica*. Patterns of TE expression in natural and synthetic *A. suecica* show that allopolyploidy is not accompanied by an overall up-regulation in TE expression as predicted by the “genome shock” hypothesis. **a** Heatmap of TE expression for the *A. thaliana* subgenome of *A. suecica* (dark green) synthetic *A. suecica* (cyan) and *A. thaliana* (light green). **b** Heatmap of TE expression for the *A. arenosa* subgenome of *A. suecica* (dark purple) synthetic *A. suecica* (pink) and *A. arenosa* (light purple). **c** and **d** the breakdown of TE families expressed in each cluster, with helitrons being the most abundant class on the *A. thaliana* subgenome and TEs of an unknown family being the most abundant in the *A. arenosa* subgenome.



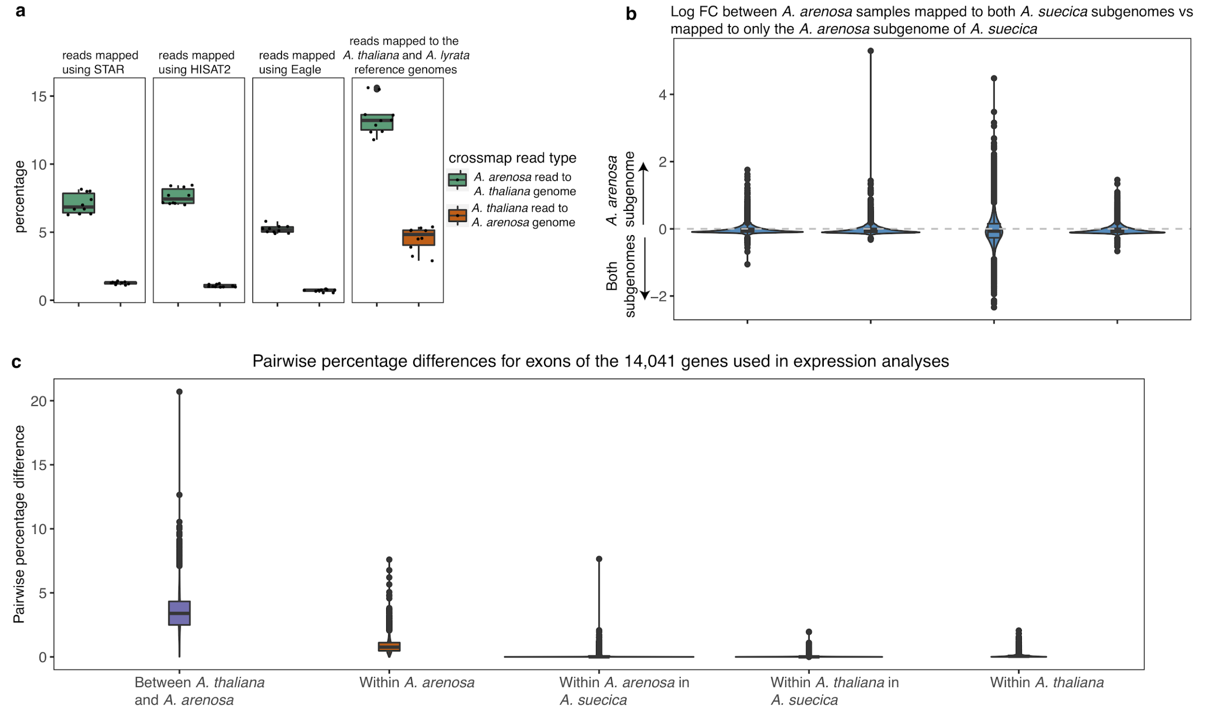
Supplementary Figure 10. Genomic distribution of TEs in the *A. suecica* genome. a Shared TEs in the population between *A. thaliana* and the *A. thaliana* subgenome of *A. suecica.* Shared TEs are likely older than private TEs and are enriched around the pericentromeric regions in the *A. thaliana* subgenome. Private TEs are enriched in the chromosomal arms for both species, where protein coding gene density is higher (Fig. 1b). **b** as in **a** but examining TEs in the population of *A. arenosa* and the *A. arenosa* part of *A. suecica*. Note the region between 5 and 10 on chromosome 2 was not included in the analysis as this region shows synteny with an unplaced contig.



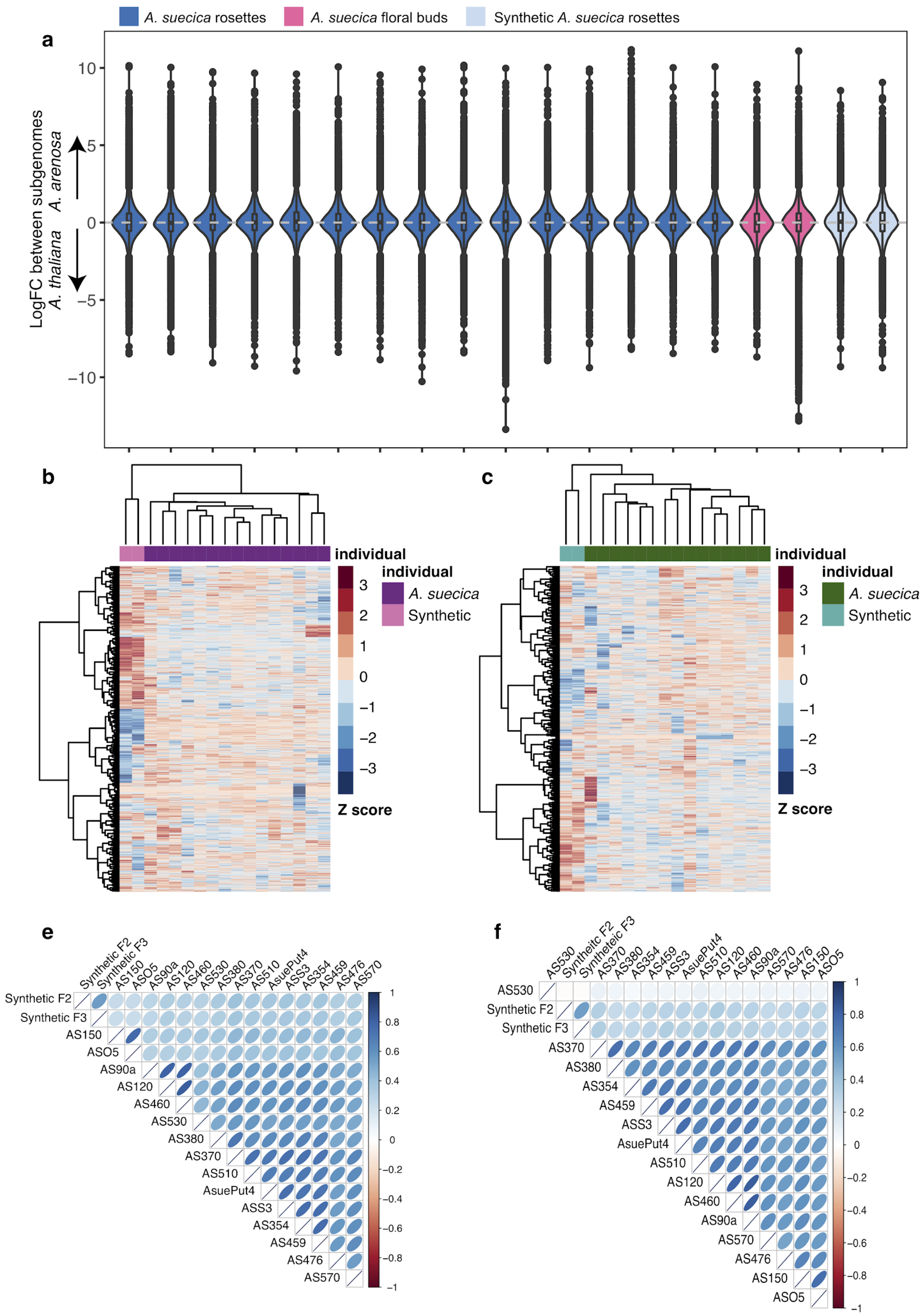
Supplementary Fig 11. Patterns of selection in *A. suecica.* a Comparison of shared variation (Nonsense SNPs, synonymous SNPs, and TEs) population frequencies in the *A. thaliana* subgenome of 15 natural *A. suecica* accessions and the closest 31 *A. thaliana* accessions. **b** Comparison of shared variation (Nonsense SNPs, synonymous SNPs, and TEs) frequencies in *A. arenosa* subgenome of 15 *A. suecica* accessions and 11 Swedish *A. arenosa* lines. Although results may be affected by the sampling and potential misidentification of the ancestral populations, the current data suggests a similar pattern on both of the subgenomes for TEs and SNPs showing a bottleneck effect. **c** Plotting quantile pairs of the population frequencies of private nonsynonymous and synonymous SNPs in *A. suecica* and ancestral populations against each other, each species shows evidence of evolution under purifying selection, since population frequency quantiles of nonsynonymous SNPs are skewed to lower values than population frequency quantiles of synonymous SNPS.



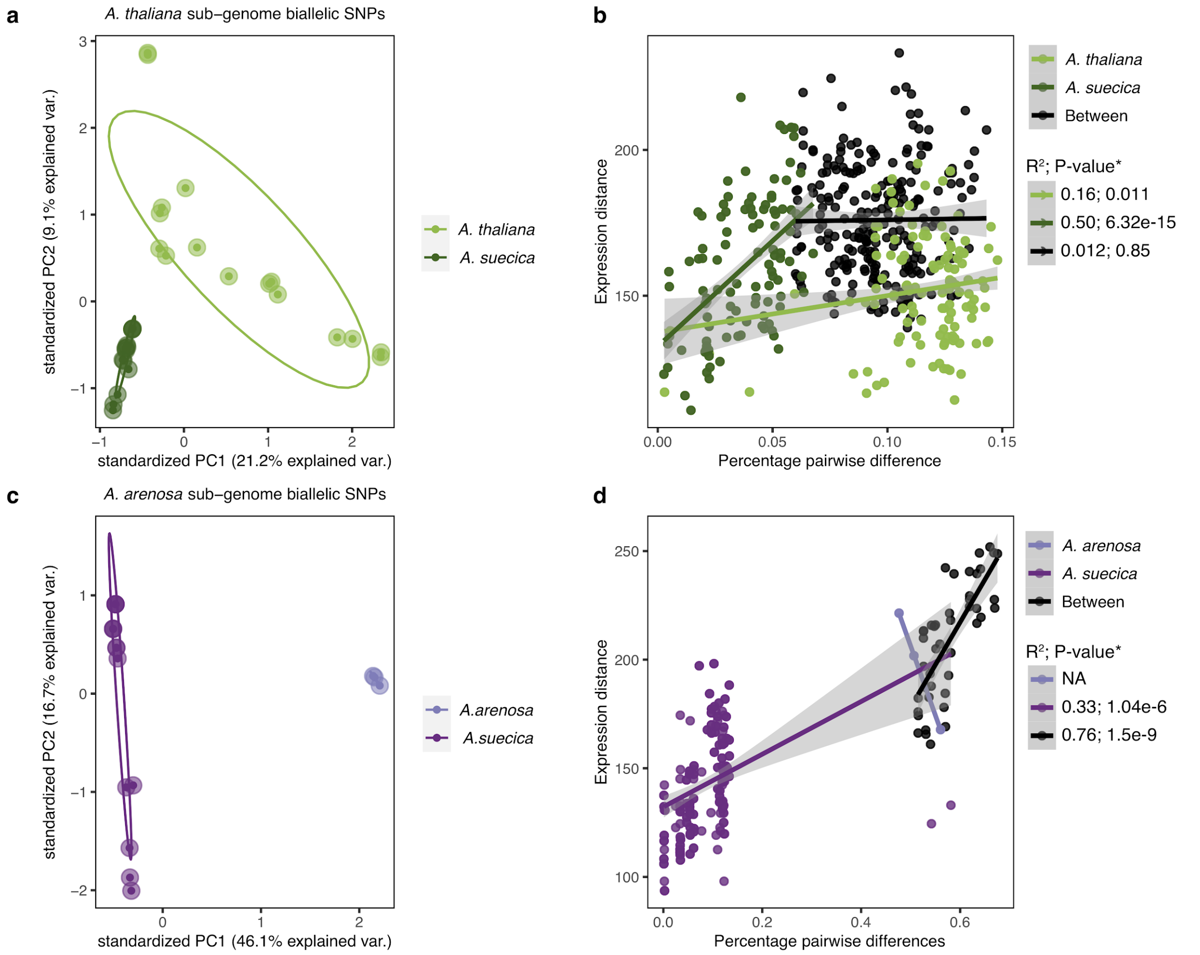
Supplementary Figure 12. Population frequencies of presence-absence calls for TEs that have mobilized between the subgenomes in *A. suecica.* a TEs ancestrally from *A. arenosa* that are present in the *A. thaliana* subgenome of *A. suecica* and **b** TEs ancestrally from *A. thaliana* that are present in the *A. arenosa* subgenome of *A. suecica*.



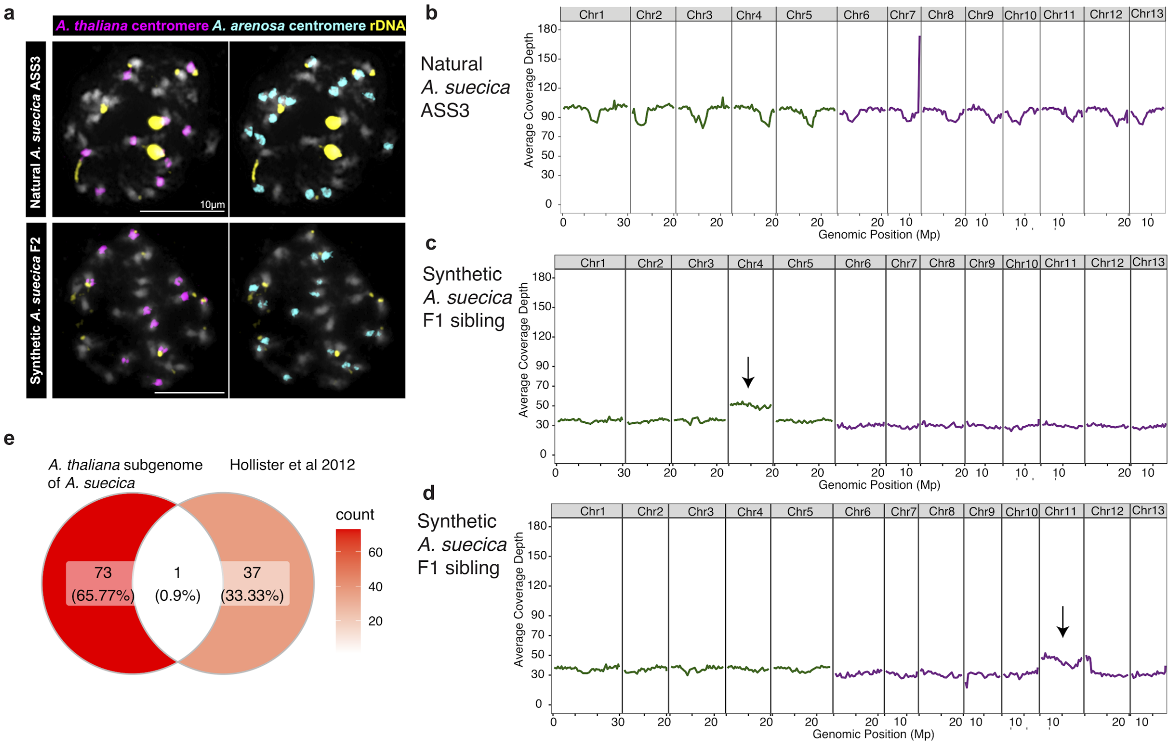
Supplementary Figure 13. Cross-mapping in RNA-seq. **a** Boxplots of cross-mapping reads. This was examined by mixing reads in-silico between *A. thaliana* and *A. arenosa*. On average ~6% of *A. arenosa* reads map to *A. thaliana* subgenome instead of the *A. arenosa* subgenome, and ~1% vice versa. Mapping these reads to the combined reference genomes of *A. thaliana* and *A. lyrata* (boxplot 4 in **a**) shows that reads map more precisely to the *A. suecica* reference and that cross-mapping is not due to unreported homeologous exchange. **b** LogFC of log2 CPM read counts for *A. arenosa* (CPM of *A. arenosa* subgenome genes when reads are mapped only to *A. arenosa* subgenome of *A. suecica*/CPM of *A. arenosa* subgenome genes when reads are mapped to the full genome) show only a small effect of mapping strategy to estimate gene expression on the *A. arenosa* subgenome. **c** Pairwise percentage differences (π) for each group measured for the exons of the 14,041 genes in the expression analysis. High levels of π in *A. arenosa* overlaps with the distribution of π between *A. thaliana* and *A. arenosa.* This explains why there is more cross-mapping for *A. arenosa* than for *A. thaliana* in **a** Importantly, lower π within *A. suecica* for both subgenomes means that measurements for subgenome dominance are not biased by cross-mapping, as we expect less cross-mapping since the distribution of π overlaps less with π between *A. thaliana* and *A. arenosa*.



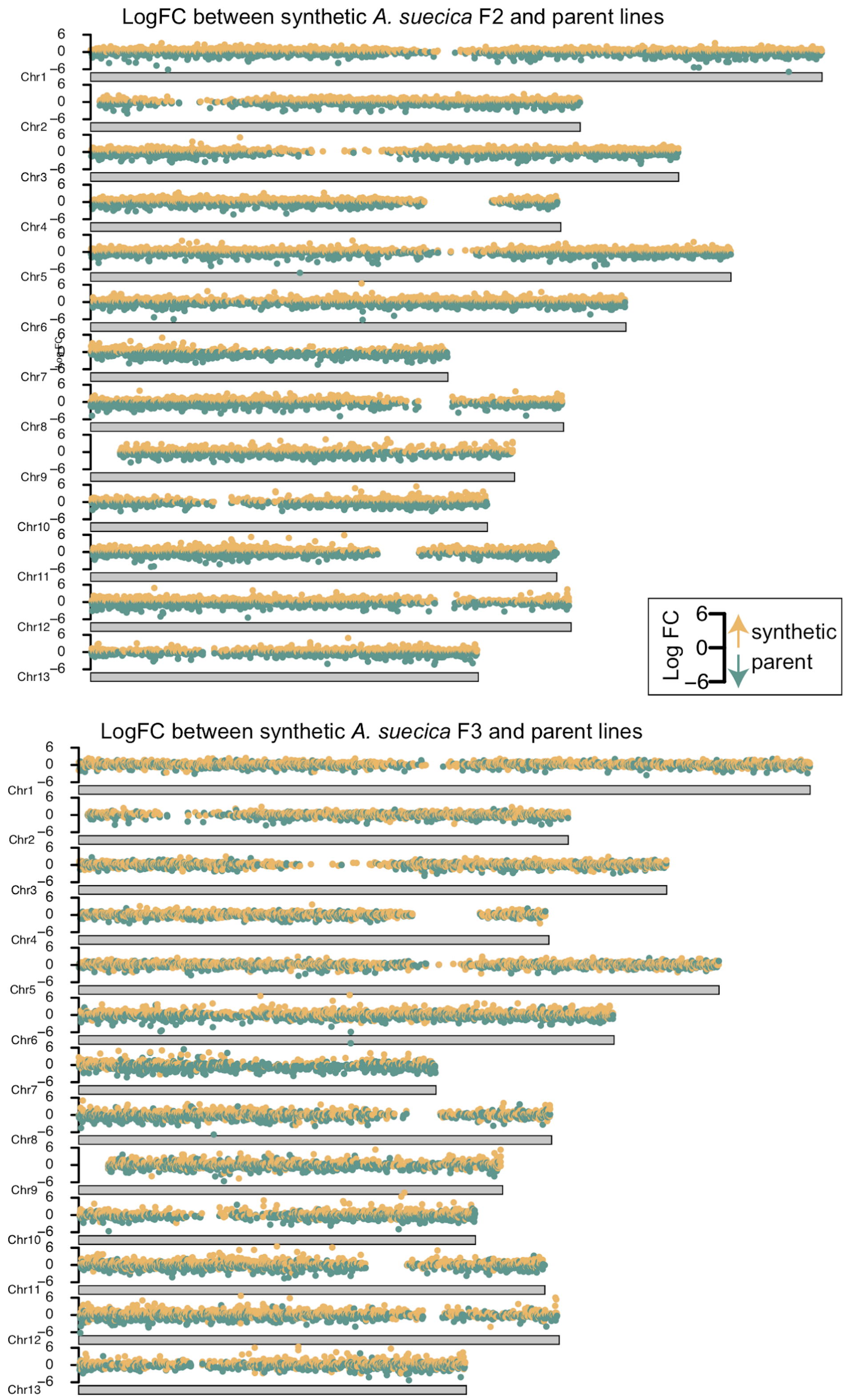
Supplemental figure 14. Expression differences between subgenomes in natural and synthetic *A. suecica*. a The distribution of expression differences across homeologous gene pairs in natural and synthetic *A. suecica*. **b** A heatmap of expression for genes in the top 5% biased toward the *A. arenosa* subgenome. The gene must be in the 5% quantile for at least 1 accession. **c** The same as in **b** but for the *A. thaliana* subgenome. Correlations of log fold change for genes in the tails of the distribution (top 5% quantile) for the *A. arenosa* subgenome **d** and the *A. thaliana* subgenome **e**



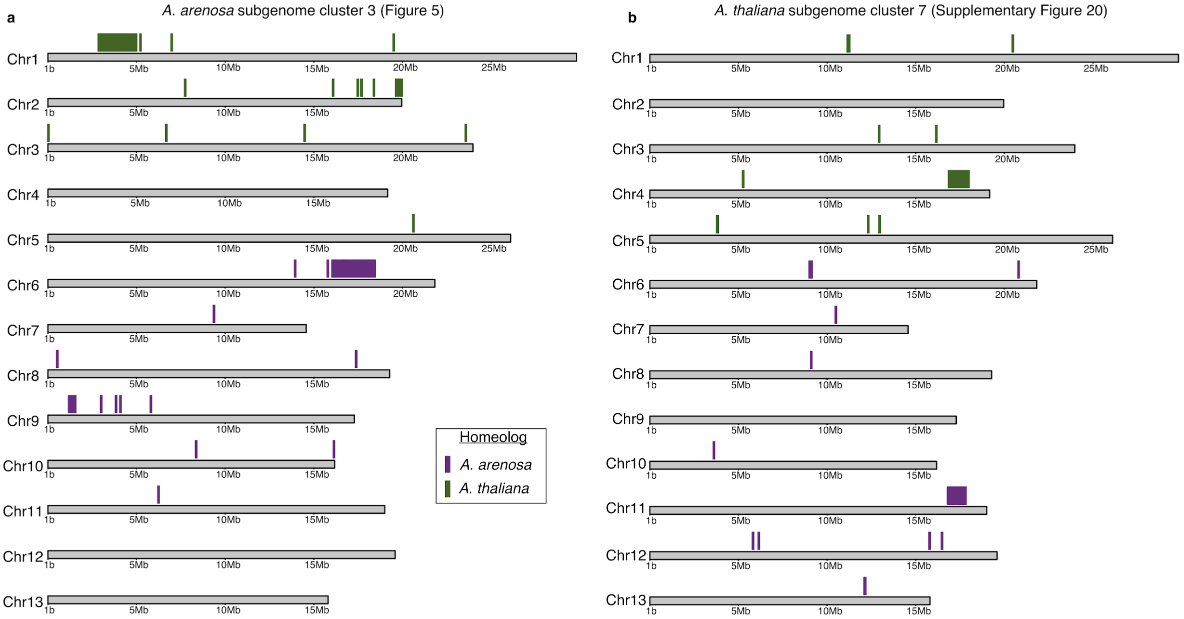
Supplementary Figure 15. Comparison of genetic and expression distance. **a** PCA plot of biallelic SNPs in the population of *A. thaliana* and *A. suecica* for the *A. thaliana* subgenome of *A. suecica* (N=345,075 biallelic SNPs), of the analyzed 13,647 genes in gene expression in addition to 500bp up and downstream of each gene sequence **b** Correlation of π (pairwise genetic differences) and expression distance (i.e. euclidean distance) for 14,041 genes (*=Bootstrapped 1000 times). **c** PCA plot of biallelic SNPs in the population of *A. arenosa* (N.B. we had DNA sequencing for only 3 of the 4 accessions used in the expression analysis) and *A. suecica* for the *A. arenosa* subgenome of *A. suecica* (N= 1,761,708 biallelic SNPs), of the analyzed 14,041 genes in gene expression in addition to 500bp up and downstream of each gene sequence **d** Correlation of Pi (pairwise genetic differences for mapped genomic regions) and expression distance (i.e. euclidean distance) for 14,041 genes (*=Bootstrapped 1000 times). *A. arenosa* was too few samples to give reliable correlations and therefore is NA. Grey bars represent the 95 confidence intervals.



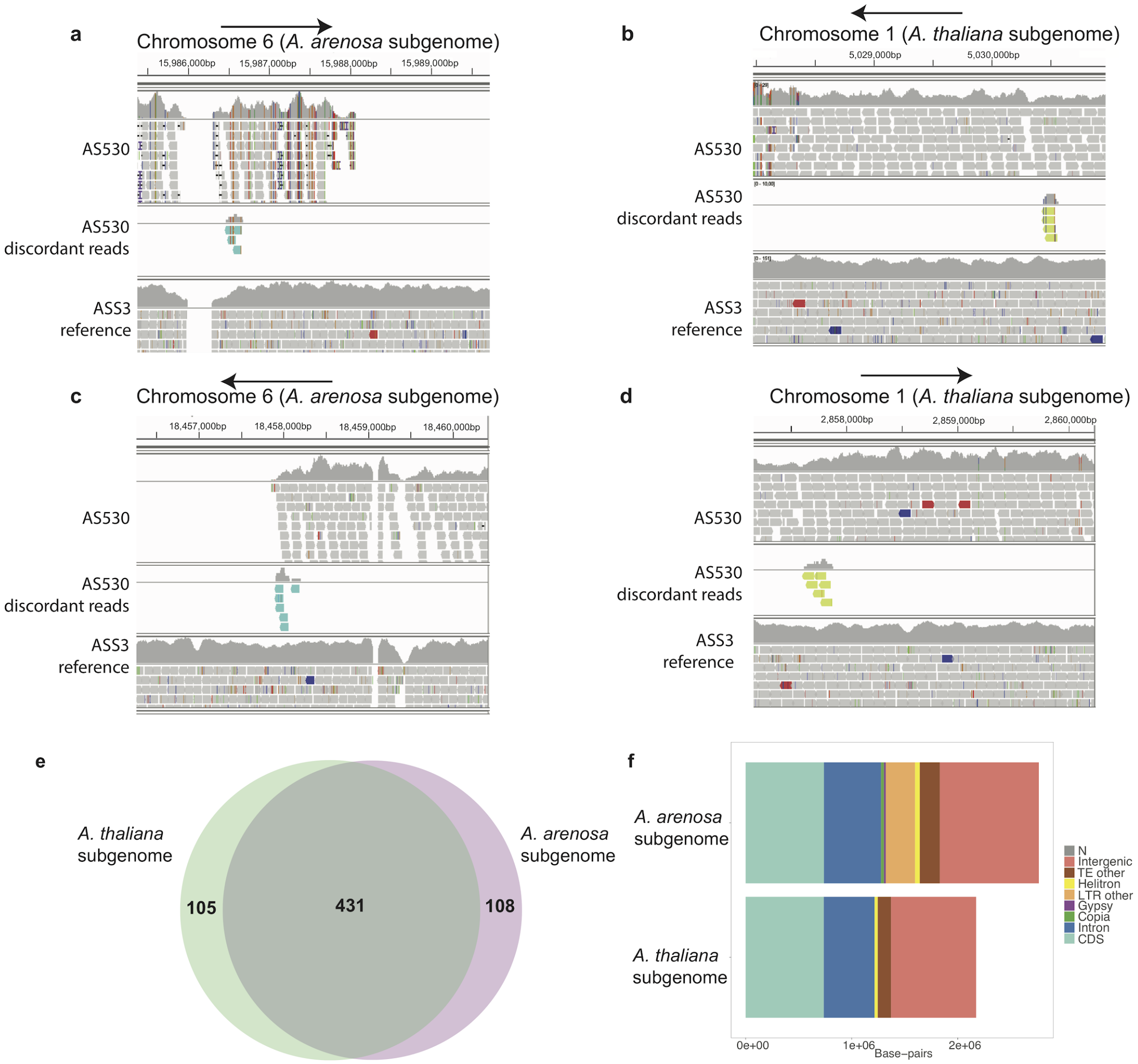
Supplementary Figure 16. **Aneuploidy is frequent in synthetic *A. suecica***. **a** Comparison of FISH analyses of the reference natural *A. suecica* “ASS3 ”and synthetic *A. suecica.* Synthetic *A. suecica* shows aneuploidy in both subgenomes in the F2 generation (gain of one chromosome on the *A. thaliana* subgenome (N=11) and loss of one chromosome on the *A. arenosa* subgenome (N=15)). Natural *A. suecica* shows a stable karyotype **b** DNA sequencing coverage in the reference natural *A. suecica* accession “ASS3” **c** and **d** DNA sequencing coverage in siblings of F1 synthetic *A. suecica* show different cases of aneuploidy (indicated with arrow) in synthetic *A. suecica*, chromosome 4 in **c** and chromosome 11 in **d** **e** overlap of genes involved in cell division from figure 5e and genes previously shown to play a role in the adaptation to autopolyploidy in *A. arenosa*[121]. The little overlap in genes between *A. suecica* and *A. arenosa* highlights that successful meiosis in polyploids is likely a complex trait.



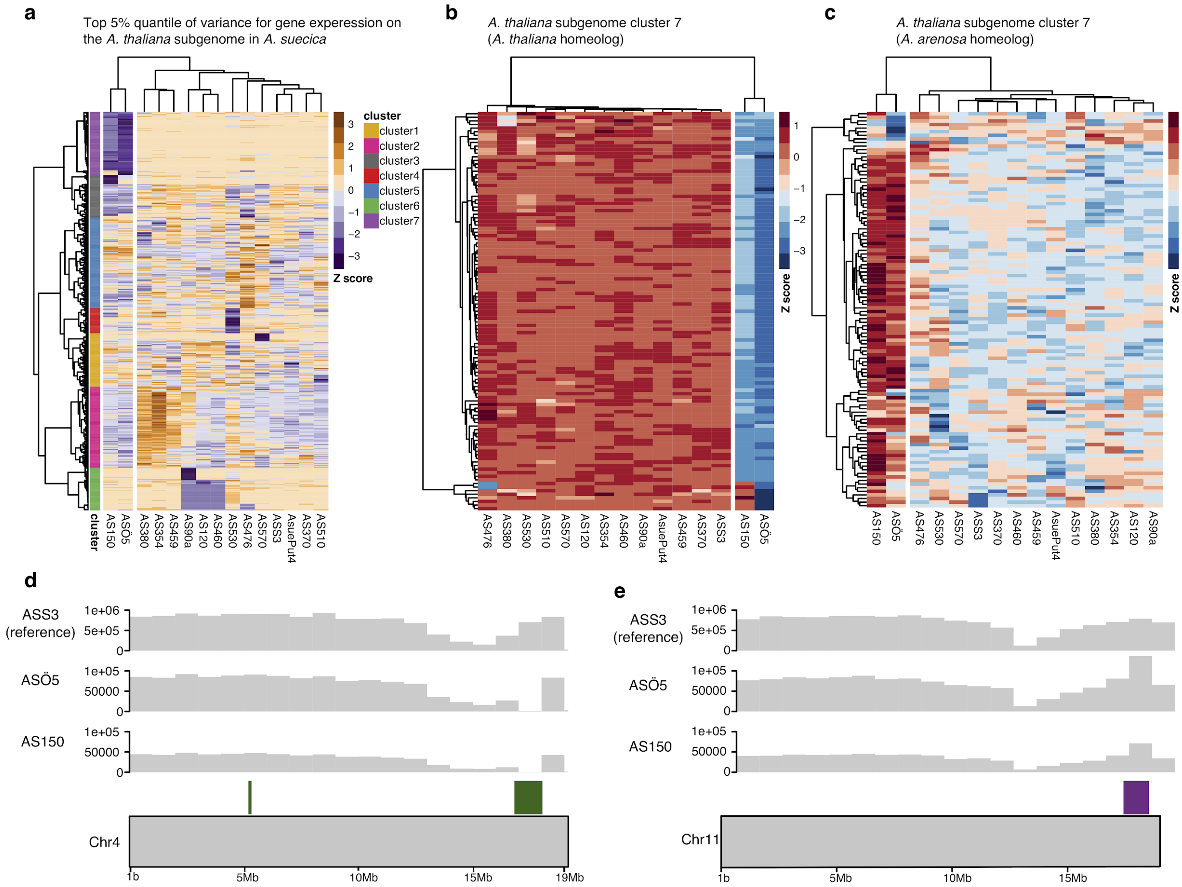
Supplementary Figure 17. No aneuploidy in synthetic *A. suecica* lines used for RNA seq based on log fold change to parent lines. Log fold change for gene expression in **a** the 2nd and **b** the 3rd generation of synthetic *A. suecica* compared to the parent lines. No clear signal of aneuploidy (i.e. an elevated increase in expression for a full chromosome) is evident.



Supplementary Figure 18 Genomic locations of genes investigated for HE signatures in *A. suecica*. a Genes in cluster 3 for Figure 5 in AS530 and **b** Genes in cluster 7 from Figure 18 in AS150 and ASÖ5

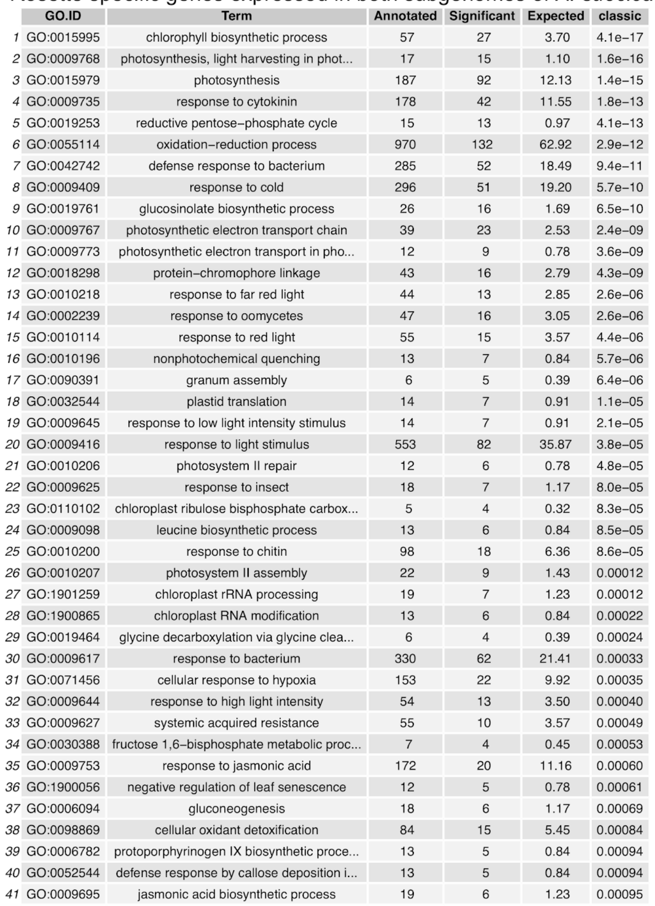


Supplementary Figure 19 Discordant read analysis supports HE in *A. suecica* a IGV screen grab of reads mapped to the beginning of the likely HE event in chromosome 6 (at ~ 15.9Mb) before coverage depth decreases to 0 in "AS530". Arrows point to the direction of the break along the chromosome. Discordant read pairs (cyan) map between the *A. arenosa* subgenome on chromosome 6 and the read pair (green) maps to the homeologous chromosome 1 on the *A. thaliana* subgenome (at ~5Mb) in **b**. The end of the likely HE event in chromosome 6 (at ~18.4Mb). Discordant reads (cyan) map between the *A. arenosa* subgenome in **c** and the read pair (green) maps to chromosome 1 (at ~2.8Mb) on the *A. thaliana* subgenome in **d**. **e** Gene counts between the syntenic regions. 431 have a 1:1 relationship, 108 genes are specific to the *A. arenosa* subgenome in this region and 105 genes are specific to the *A. thaliana* subgenome. **f** Composition of the syntenic regions between the two subgenomes



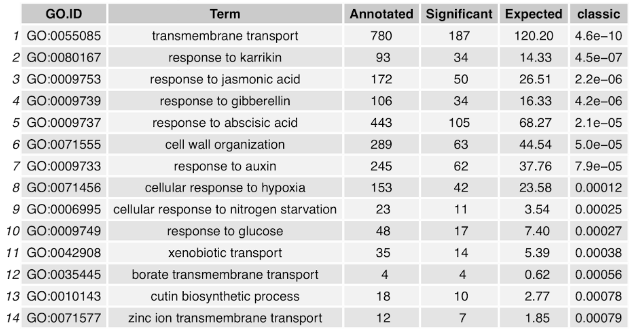
Supplementary Figure 20. Homeologous exchange contributes to expression variance within *A. suecica* on the *A. thaliana* subgenome. a Taking the top 5% quantiles (N=702) for variation in gene expression for the *A. thaliana* subgenome we find a large cluster 7 (N=111) where the two outlier accessions in our PCA ("AS150" and "ASÖ5") are expressing these genes differently to the rest of the population. **b** Homeologous genes of this cluster on the *A. thaliana* subgenome of *A. suecica* show that these genes are not expressed in these two accessions while **c** shows the opposite pattern and are higher expressed in "AS150" and "ASÖ5" compared to the rest of the population. **d** 101 of the 111 genes in cluster 7 are located on chromosome 4 in close proximity to each other on the *A. thaliana* subgenome of the *A. suecica* reference genome and appear to be deleted in "AS5Ö5" and "AS150" as they do not have DNA sequencing coverage. The *A. arenosa* subgenome homeologs (located on chromosome 11) have twice the DNA coverage, suggesting they are duplicated, in agreement with the expectations of HE event.

a Rosette specific genes expressed in both subgenomes of *A. suecica*



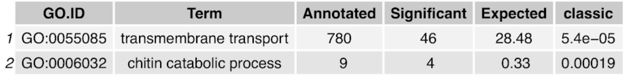
	GO.ID	Term	Annotated	Significant	Expected	classic
1	GO:0015995	chlorophyll biosynthetic process	57	27	3.70	4.1e−17
2	GO:0009768	photosynthesis, light harvesting in phot...	17	15	1.10	1.6e−16
3	GO:0015979	photosynthesis	187	92	12.13	1.4e−15
4	GO:0009735	response to cytokinin	178	42	11.55	1.8e−13
5	GO:0019253	reductive pentose−phosphate cycle	15	13	0.97	4.1e−13
6	GO:0055114	oxidation−reduction process	970	132	62.92	2.9e−12
7	GO:0042742	defense response to bacterium	285	52	18.49	9.4e−11
8	GO:0009409	response to cold	296	51	19.20	5.7e−10
9	GO:0019761	glucosinolate biosynthetic process	26	16	1.69	6.5e−10
10	GO:0009767	photosynthetic electron transport chain	39	23	2.53	2.4e−09
11	GO:0009773	photosynthetic electron transport in pho...	12	9	0.78	3.6e−09
12	GO:0018298	protein−chromophore linkage	43	16	2.79	4.3e−09
13	GO:0010218	response to far red light	44	13	2.85	2.6e−06
14	GO:0002239	response to oomycetes	47	16	3.05	2.6e−06
15	GO:0010114	response to red light	55	15	3.57	4.4e−06
16	GO:0010196	nonphotochemical quenching	13	7	0.84	5.7e−06
17	GO:0090391	granum assembly	6	5	0.39	6.4e−06
18	GO:0032544	plastid translation	14	7	0.91	1.1e−05
19	GO:0009645	response to low light intensity stimulus	14	7	0.91	2.1e−05
20	GO:0009416	response to light stimulus	553	82	35.87	3.8e−05
21	GO:0010206	photosystem II repair	12	6	0.78	4.8e−05
22	GO:0009625	response to insect	18	7	1.17	8.0e−05
23	GO:0110102	chloroplast ribulose bisphosphate carbox...	5	4	0.32	8.3e−05
24	GO:0009098	leucine biosynthetic process	13	6	0.84	8.5e−05
25	GO:0010200	response to chitin	98	18	6.36	8.6e−05
26	GO:0010207	photosystem II assembly	22	9	1.43	0.00012
27	GO:1901259	chloroplast rRNA processing	19	7	1.23	0.00012
28	GO:1900865	chloroplast RNA modification	13	6	0.84	0.00022
29	GO:0019464	glycine decarboxylation via glycine clea...	6	4	0.39	0.00024
30	GO:0009617	response to bacterium	330	62	21.41	0.00033
31	GO:0071456	cellular response to hypoxia	153	22	9.92	0.00035
32	GO:0009644	response to high light intensity	54	13	3.50	0.00040
33	GO:0009627	systemic acquired resistance	55	10	3.57	0.00049
34	GO:0030388	fructose 1,6−bisphosphate metabolic proc...	7	4	0.45	0.00053
35	GO:0009753	response to jasmonic acid	172	20	11.16	0.00060
36	GO:1900056	negative regulation of leaf senescence	12	5	0.78	0.00061
37	GO:0006094	gluconeogenesis	18	6	1.17	0.00069
38	GO:0098869	cellular oxidant detoxification	84	15	5.45	0.00084
39	GO:0006782	protoporphyrinogen IX biosynthetic proce...	13	5	0.84	0.00094
40	GO:0052544	defense response by callose deposition i...	13	5	0.84	0.00094
41	GO:0009695	jasmonic acid biosynthetic process	19	6	1.23	0.00095

b Floral bud specific genes expressed in both subgenomes of *A. suecica*



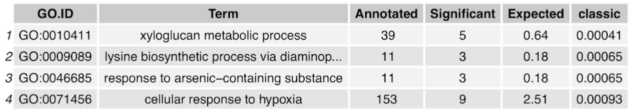
	GO.ID	Term	Annotated	Significant	Expected	classic
1	GO:0055085	transmembrane transport	780	187	120.20	4.6e−10
2	GO:0080167	response to karrikin	93	34	14.33	4.5e−07
3	GO:0009753	response to jasmonic acid	172	50	26.51	2.2e−06
4	GO:0009739	response to gibberellin	106	34	16.33	4.2e−06
5	GO:0009737	response to abscisic acid	443	105	68.27	2.1e−05
6	GO:0071555	cell wall organization	289	63	44.54	5.0e−05
7	GO:0009733	response to auxin	245	62	37.76	7.9e−05
8	GO:0071456	cellular response to hypoxia	153	42	23.58	0.00012
9	GO:0006995	cellular response to nitrogen starvation	23	11	3.54	0.00025
10	GO:0009749	response to glucose	48	17	7.40	0.00027
11	GO:0042908	xenobiotic transport	35	14	5.39	0.00038
12	GO:0035445	borate transmembrane transport	4	4	0.62	0.00056
13	GO:0010143	cutin biosynthetic process	18	10	2.77	0.00078
14	GO:0071577	zinc ion transmembrane transport	12	7	1.85	0.00079

c Floral bud specific genes expressed biased towards *A. arenosa*



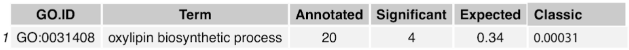
	GO.ID	Term	Annotated	Significant	Expected	classic
1	GO:0055085	transmembrane transport	780	46	28.48	5.4e−05
2	GO:0006032	chitin catabolic process	9	4	0.33	0.00019

d Rosette specific genes expressed biased towards *A. arenosa*



	GO.ID	Term	Annotated	Significant	Expected	classic
1	GO:0010411	xyloglucan metabolic process	39	5	0.64	0.00041
2	GO:0009089	lysine biosynthetic process via diaminop...	11	3	0.18	0.00065
3	GO:0046685	response to arsenic−containing substance	11	3	0.18	0.00065
4	GO:0071456	cellular response to hypoxia	153	9	2.51	0.00093

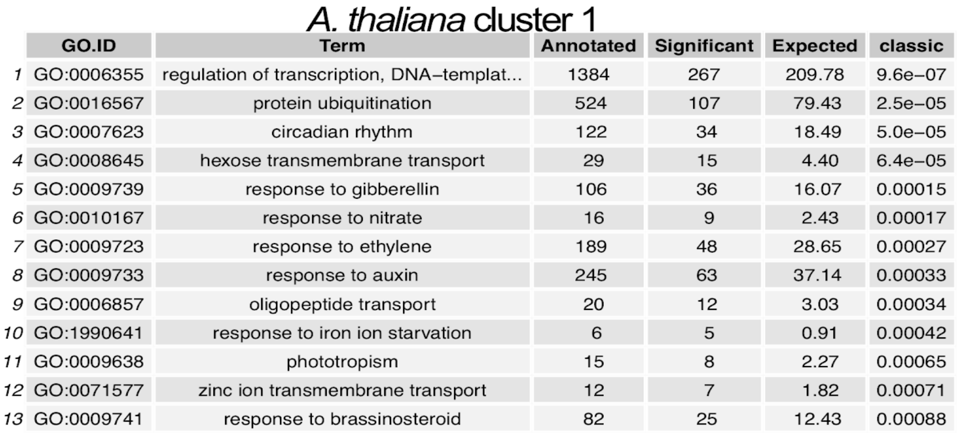
e Rosette specific genes expressed biased towards *A. thaliana*



	GO.ID	Term	Annotated	Significant	Expected	Classic
1	GO:0031408	oxylipin biosynthetic process	20	4	0.34	0.00031

Supplementary Table 1. Gene ontology (GO) analysis for gene expression comparison between whole rosettes and floral buds in *A. suecica*. No significant GO was found for genes biased towards the *A. thaliana* subgenome of *A. suecica* for floral buds.

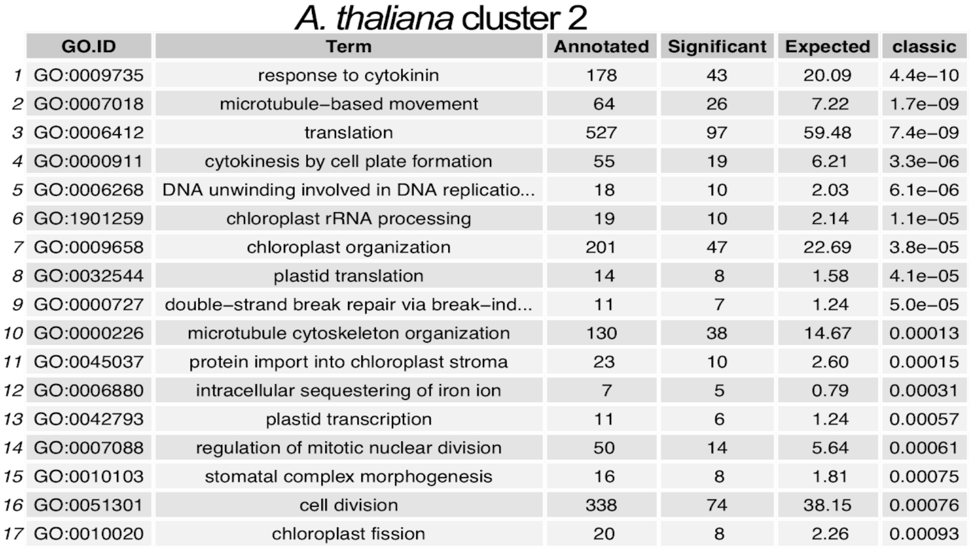
a



A. thaliana cluster 1

	GO.ID	Term	Annotated	Significant	Expected	classic
1	GO:0006355	regulation of transcription, DNA−templat...	1384	267	209.78	9.6e−07
2	GO:0016567	protein ubiquitination	524	107	79.43	2.5e−05
3	GO:0007623	circadian rhythm	122	34	18.49	5.0e−05
4	GO:0008645	hexose transmembrane transport	29	15	4.40	6.4e−05
5	GO:0009739	response to gibberellin	106	36	16.07	0.00015
6	GO:0010167	response to nitrate	16	9	2.43	0.00017
7	GO:0009723	response to ethylene	189	48	28.65	0.00027
8	GO:0009733	response to auxin	245	63	37.14	0.00033
9	GO:0006857	oligopeptide transport	20	12	3.03	0.00034
10	GO:1990641	response to iron ion starvation	6	5	0.91	0.00042
11	GO:0009638	phototropism	15	8	2.27	0.00065
12	GO:0071577	zinc ion transmembrane transport	12	7	1.82	0.00071
13	GO:0009741	response to brassinosteroid	82	25	12.43	0.00088

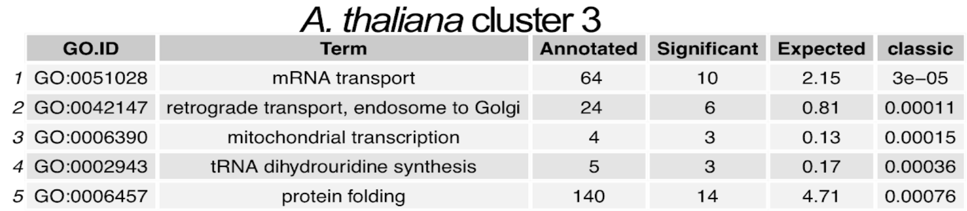
b



A. thaliana cluster 2

	GO.ID	Term	Annotated	Significant	Expected	classic
1	GO:0009735	response to cytokinin	178	43	20.09	4.4e−10
2	GO:0007018	microtubule−based movement	64	26	7.22	1.7e−09
3	GO:0006412	translation	527	97	59.48	7.4e−09
4	GO:0000911	cytokinesis by cell plate formation	55	19	6.21	3.3e−06
5	GO:0006268	DNA unwinding involved in DNA replicatio...	18	10	2.03	6.1e−06
6	GO:1901259	chloroplast rRNA processing	19	10	2.14	1.1e−05
7	GO:0009658	chloroplast organization	201	47	22.69	3.8e−05
8	GO:0032544	plastid translation	14	8	1.58	4.1e−05
9	GO:0000727	double−strand break repair via break−ind...	11	7	1.24	5.0e−05
10	GO:0000226	microtubule cytoskeleton organization	130	38	14.67	0.00013
11	GO:0045037	protein import into chloroplast stroma	23	10	2.60	0.00015
12	GO:0006880	intracellular sequestering of iron ion	7	5	0.79	0.00031
13	GO:0042793	plastid transcription	11	6	1.24	0.00057
14	GO:0007088	regulation of mitotic nuclear division	50	14	5.64	0.00061
15	GO:0010103	stomatal complex morphogenesis	16	8	1.81	0.00075
16	GO:0051301	cell division	338	74	38.15	0.00076
17	GO:0010020	chloroplast fission	20	8	2.26	0.00093

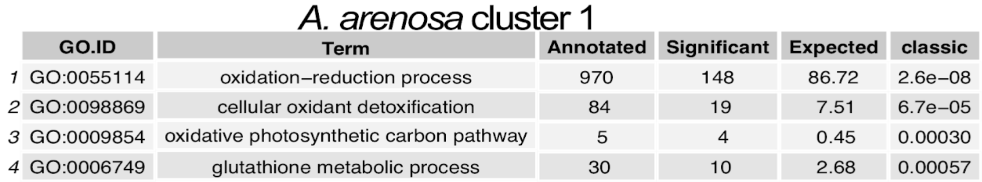
c



A. thaliana cluster 3

	GO.ID	Term	Annotated	Significant	Expected	classic
1	GO:0051028	mRNA transport	64	10	2.15	3e−05
2	GO:0042147	retrograde transport, endosome to Golgi	24	6	0.81	0.00011
3	GO:0006390	mitochondrial transcription	4	3	0.13	0.00015
4	GO:0002943	tRNA dihydrouridine synthesis	5	3	0.17	0.00036
5	GO:0006457	protein folding	140	14	4.71	0.00076

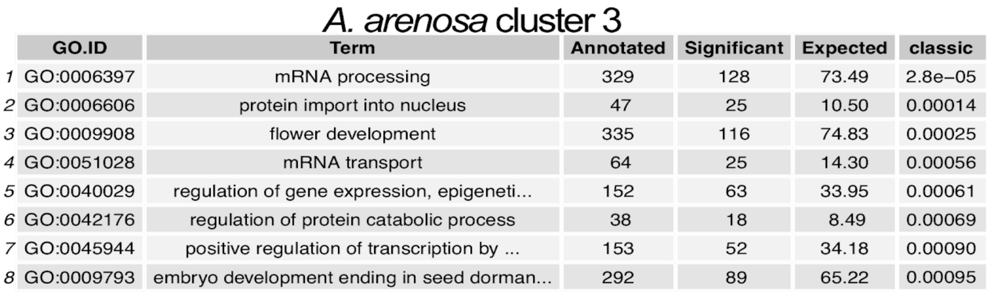
d



A. arenosa cluster 1

	GO.ID	Term	Annotated	Significant	Expected	classic
1	GO:0055114	oxidation−reduction process	970	148	86.72	2.6e−08
2	GO:0098869	cellular oxidant detoxification	84	19	7.51	6.7e−05
3	GO:0009854	oxidative photosynthetic carbon pathway	5	4	0.45	0.00030
4	GO:0006749	glutathione metabolic process	30	10	2.68	0.00057

e



A. arenosa cluster 3

	GO.ID	Term	Annotated	Significant	Expected	classic
1	GO:0006397	mRNA processing	329	128	73.49	2.8e−05
2	GO:0006606	protein import into nucleus	47	25	10.50	0.00014
3	GO:0009908	flower development	335	116	74.83	0.00025
4	GO:0051028	mRNA transport	64	25	14.30	0.00056
5	GO:0040029	regulation of gene expression, epigeneti...	152	63	33.95	0.00061
6	GO:0042176	regulation of protein catabolic process	38	18	8.49	0.00069
7	GO:0045944	positive regulation of transcription by ...	153	52	34.18	0.00090
8	GO:0009793	embryo development ending in seed dorman...	292	89	65.22	0.00095

Supplementary Table 2. List of overrepresented gene ontologies on the Fig. 5e

Materials & Methods

PacBio sequencing of *A. suecica*

We used genomic DNA from whole rosettes of one *A. suecica* (“ASS3”) accession to generate PacBio sequencing data. DNA was extracted using a modified PacBio protocol for preparing *Arabidopsis* genomic DNA for size-selected ~20kb SMRTbell libraries. Briefly, whole genomic DNA was extracted from 32g of 3-4 week old plants, grown at 16°C and subjected to a 2-day dark treatment. This generated 23 micrograms of purified genomic DNA with a fragment length of >40Kb for *A. suecica*. We assessed DNA quality with a Qubit fluorometer and a Nanodrop analysis, and ran the DNA on a gel to visualize fragmentation. Genomic libraries and single-molecule real-time (SMRT) sequence data were generated at the Functional Genomics Center Zurich (FGCZ), in Switzerland. The Pacbio RSII instrument was used with P6/C4 chemistry and an average movie length of 6 hours. A total of 12 SMRT cells were processed generating 16.3Gb of DNA bases with an N50 read length of 20 Kbp and median read length of 14 Kbp. Using the same genomic library, an additional 3.3 Gbp of data was generated by a Pacbio Sequel instrument at the Vienna Biocenter Core Facilities (VBCF), in Austria, with a median read length of 10Kbp.

A. suecica genome assembly

To generate the *A. suecica* assembly we first used FALCON[127] (version 0.3.0) with a length cutoff for seed reads set to 1 Kb in size. The assembly produced 828 contigs with an N50 of 5.81 Mb and a total assembly size of 271 Mb. Additionally, we generated a Canu[128] (v.1.3.0) assembly using default settings, which resulted in 260 contigs with an N50 of 6.65 Mb and a total assembly size of 267 Mb. Then we merged the two assemblies using the software quickmerge[129]. The resulting merged assembly consisted of 929 contigs with an N50 of 9.02 Mb and a total draft assembly size of 276 Mb. We polished the assembly using Arrow[130] (smrtlink release 5.0.0.6792) and Pilon (version 1.22). For Pilon[131], 100bp (with PCR duplicates removed), and a second PCR-free 250bp, Illumina paired end reads were used that had been generated from the reference *A. suecica* accession “ASS3”.

Pacbio sequencing of *A. arenosa*

A natural Swedish autotetraploid *A. arenosa* accession "Aa4" was inbred in a lab for two generations in order to reduce heterozygosity. We extracted whole genomic DNA from 64g of three week old plants in the same way as described for *A. suecica* (above), generating 50 µg of purified genomic DNA with a fragment sizes longer than 40 Kb in length. The *A. arenosa* genomic libraries and SMRT sequence data were generated at the Vienna Biocenter Core Facilities (VBCF), in Austria. A Pacbio Sequel instrument was used to generate a total of 22 Gbp of data from five SMRT cells, with an N50 of 13 Kbp and median read length 10 Kbp. In addition, two runs of Oxford Nanopore sequencing were carried out at the VBCF producing 750 Mbp in 180,000 reads (median 5 Kbp and 2.6 Kbp; N50 8.7 and 6.7 Kbp, respectively).

Assembly of autotetraploid *A. arenosa*

We assembled a draft contig assembly for the autotetraploid *A. arenosa* accession "Aa4" using FALCON (version 0.3.0) as for *A. suecica*. The assembly produced 3,629 contigs with an N50 of 331 Kb, maximum contig size of 2.5 Mb and a total assembly size of 461 Mb. The assembly size is greater than the calculated haploid size of 330 Mb using FACs (see Supplementary Figure 2) probably because of the high levels of heterozygosity in *A. arenosa*. The resulting assembly was polished as described for *A. suecica*.

HiC tissue fixation and library preparation

To generate physical scaffolds for the *A. suecica* assembly we generated proximity-ligation HiC sequencing data. We collected approximately 0.5 gram of tissue from 3-week old seedlings of the same reference *A. suecica* accession. Freshly collected plant tissue was fixed in 1% formaldehyde. Cross-linking was stopped by the addition of 0.15 M Glycine. The fixed tissue was ground to a powder in liquid nitrogen and suspended in 10 ml of nuclei isolation buffer. Nuclei was digested by adding 50 U DpnII and the digested chromatin was blunt-ended by incubation with 25 µL of 0.4 mM biotin-14-dCTP and 40 U of Klenow enzyme, as described in [ref]. 20 U of T4 DNA ligase was then added to start proximity ligation. The extracted DNA was sheared by sonication with a Covaris S220 to produce 250-500bp fragments. This was followed by size fractionation using AMPure XP beads. Biotin was then removed from unligated ends. DNA fragments were blunt-end repaired and adaptors were ligated to the DNA products following the NEBNext Ultra II RNA Library Prep Kit for Illumina.

To analyse structural rearrangements we collected tissue for 1 other natural *A. suecica* “AS530”, 1 *A. thaliana* accession ”6978”, 1 *A. arenosa* “Aa6” and 1 synthetic *A. suecica* (F3). Each sample had two replicates. We collected tissue and prepared libraries in the same manner as described above. 125bp paired-end Illumina reads were mapped using HiCUP[132] (version 0.6.1).

Reference-guided scaffolding of the *A. suecica* genome with *LACHESIS*

We sequenced 207 million pairs of 125bp paired-end Illumina reads from the HiC library of the reference accession "ASS3". We mapped reads using HiCUP (version 0.6.1) to the draft *A. suecica* contig assembly. This resulted in ~137 million read pairs with a unique alignment.

Setting an assembly threshold of >= 1 Kb in size, contigs of the draft *A. suecica* assembly were first assigned to the *A. thaliana* or *A. arenosa* subgenome. To do this, we used nucmer from the software MUMmer[133] (version 3.23) to perform whole-genome alignments. We aligned the draft *A. suecica* assembly to the *A. thaliana* TAIR10 reference and to our *A. arenosa* draft contig assembly, simultaneously. We used the MUMer command dnadiff to produce 1-to-1 alignments. As the subgenomes are only ~86% identical, the majority of contigs could be conclusively assigned to either subgenome by examining how similar the alignments were. Contigs that could not be assigned to a subgenome based on percentage identity were examined manually, and the length of the alignment was used to determine subgenome assignment.

Finally, we used the software LACHESIS[134] (version 1.0.0) to scaffold our draft assembly, using the reference genomes of *A. thaliana* and *A. lyrata as* a guide to assist with scaffolding the contigs (we used *A. lyrata* here instead of our draft *A. arenosa* contig assembly, as *A. lyrata* is a chromosome-level assembly). This produced a 13-scaffold chromosome-level assembly for *A. suecica*.

Construction of the *A. suecica* genetic map

We crossed natural *A. suecica* accession "AS150" with the reference accession "ASS3". The cross was uni-directional with "AS150" as the maternal and "ASS3" as the paternal plant. F1 plants were grown, and F2 seeds were collected, from which we grew and collected 192 F2 plants. We multiplexed the samples on 96 well plates using 75bp paired end reads and generated data of 1-2x coverage per sample. Samples were mapped to the repeat-masked scaffolds of the reference *A. suecica* genome using BWA-MEM[135] (version 0.7.15). Samtools[136] (version 0.1.19) was used to filter reads for proper pairs and a minimum mapping quality of 5 (-F 256 -f 3 -q 5). We called variants directly from samtools mpileup output on the sequenced F2 individuals at known biallelic sites between the two accessions used to generate the cross (a total of 590,537 SNPs). We required sites to have non-zero coverage in a minimum of 20 individuals and filtered SNPs to have frequency between 0.45-0.55 in our F2 population (as the expectation is 50:50),. We removed F2 individuals that did not have genotype calls for more than 90% of the data. This resulted in 183 individuals with genotype calls for 334,257 SNPs.

Since sequencing coverage for the F2s was low this meant we had a low probability of calling heterozygous SNPs, and a higher probability of calling a SNP as homozygous. Therefore, we applied a Hidden Markov Model implemented in R package HMM[137] to classify SNPs as homozygous or heterozygous for each of our F2 lines. We then divided the genome into 500Kb non-overlapping windows, and classified each window as homozygous (here 0 or 1, for the reference or alternate SNP) or heterozygous (here 0.5). If the frequency of 1, 0 or 0.5 represented more than 50% of the SNPs in a given window, and exceeded missing calls (NA), the window was designated as 1, 0 or 0.5 (otherwise it was NA). This was done per chromosome and the resulting file for each chromosome and their markers were processed in the R package qtl[138], in order to generate a genetic map. Markers genotyped in less than 100 F2s were excluded from the analysis. Linkage groups were assigned with a minimum LOD score of 8 and a maximum recombination fraction of 0.35. Each chromosome was assigned to one linkage group. We defined the final marker order by the best LOD score and the lowest number of crossover events.

Notably, the assistance of a genetic map corrected the erroneous placement of a contig at the beginning of chromosome 1 of the *A. arenosa* subgenome. The misplaced contig was relocated from chromosome 1 to the pericentromeric region of chromosome 2 of the *A. arenosa* subgenome in *A. suecica*. This error was a result of a mis-assembly of chromosome 1 in the *A. lyrata* reference, as was previously pointed out [77]. Also of note, chromosome 2 of the *A. thaliana* subgenome of *A. suecica* was previously shown to be largely devoid of intraspecific variation, thus we had sparse marker information for this chromosome in the genetic map. Therefore, this chromosome-scale scaffold was largely assembled by the manual inspection of 3D-proximity information based on our HiC sequencing and reviewing contig order using the software Juicebox[139].

Gene prediction and annotation of the *A. suecica* genome

We combined *de novo* and evidence-based approaches to predict protein coding genes. For *de novo* prediction, we trained AUGUSTUS[140] on the set of conserved single copy genes using BUSCO[141] separately on *A. thaliana* and *A. arenosa* subgenomes of *A. suecica*. The evidence-based approach included both homology to the protein sequences of the ancestral species and the transcriptome of *A. suecica*. We aligned the peptide sequences from TAIR10 *A. thaliana* assembly to the *A. thaliana* subgenome of *A. suecica*, while the peptides from *A. lyrata* from the second version of *A. lyrata* annotation[142] (Alyrata_384_v2.1) were aligned to the *A. arenosa* subgenome of *A. suecica* using GenomeThreader[143] (1.7.0). We mapped the RNAseq reads from the reference accession of *A. suecica* (ASS3) from the rosettes and flower buds tissues (see above) to the reference genome using tophat[144] and generated intron hints from the split reads using bam2hints extension of AUGUSTUS. We split the alignment into *A. thaliana* and *A. arenosa* subgenomes and assembled the transcriptome of *A. suecica* for each subgenome separately in the genome-guided mode with Trinity[145] (2.6.6). Separately for each of the subgenomes, we filtered the assembled transcripts using tpm cutoff set to 1, collapsed similar transcripts using CD-HIT[146,147] with sequence identity set to 90 percent, and chose the longest open reading frame from the six-frame translation. We then aligned the proteins from *A. thaliana* and *A. arenosa* parts of *A. suecica* to the corresponding subgenomes using GenomeThreader (1.7.0). We ran AUGUSTUS using retrained parameters from BUSCO and merged hints from all three sources, these being: (1) intron hints from *A. suecica* RNAseq, (2) homology hints from ancestral proteins and (3) hints from *A. suecica* proteins.

RepeatModeler[148] (version 1.0.11) was used in order to build a *de novo* TE consensus library for *A. suecica* and identify repetitive elements based on the genome sequence. Genome locations for the identified TE repeats were determined by using RepeatMasker[149] (version 4.0.7) and filtered for full length matches using a code described in Bailly-Bechet et. al[150]. Helitrons are the most abundant TE family in both subgenomes (Supplementary Fig. 7).

Synthetic *A. suecica* lines

To generate synthetic *A. suecica* we crossed a natural tetraploid *A. thaliana* accession (6978 aka "Wa-1") to a natural Swedish autotetraploid *A. arenosa* ("Aa4") accession. Similar to the natural *A. suecica*, *A. thaliana* was the maternal and *A. arenosa* was the paternal plant in this cross. Crosses in the opposite direction were unsuccessful. We managed to obtain very few F1 hybrid plants, which after one round of selfing set higher levels of seed formation. The resulting synthetic line was able to self-fertilize. F2 seeds were descended from a common F1 and were similar to natural *A. suecica* in appearance. We further continued the synthetic line to F3 (selfed 3rd generation).

Synteny analysis

We performed all-against-all BLASTP search using CDS sequences for the reference *A. suecica* genome and the ancestral genomes, *A. thaliana* and *A. lyrata* (here the closest substitute reference genome for *A. arenosa*, with annotation). We used the SynMap tool[151] from the online CoGe portal[152]. We examined synteny using the default parameters for DAGChainer (maximum distance between two matches = 20 genes; minimum number of aligned pairs = 5 genes).

Estimating copy number of rDNA repeats using short DNA reads

To measure copy number of 45S rRNA repeats in our populations of different species, we aligned short DNA reads to a single reference 45S consensus sequence of *A. thaliana*[153]. An *A. arenosa* 45S rRNA consensus sequence was constructed by finding the best hit using BLAST in our draft *A. arenosa* contig assembly. This hit matched position 1571-8232 bp of the *A. thaliana* consensus sequence, was 6,647 bp in length and is 97% identical to the *A. thaliana* 45s rRNA consensus sequence. The aligned regions of these two 45S rRNA consensus sequences, determined by BLAST, were used in copy number estimates, to ensure that the size of the sequences were equal. The relative increase in sequence coverage of these loci, when compared to the mean coverage for the reference genome, was used to estimate copy number.

Plant material for RNA sequencing

Transcriptomic data generated in this study included 15 accessions of *A. suecica*, 16 accessions of *A. thaliana*, 4 accessions of *A. arenosa* and 2 generations of an artificial *A. suecica* line (the 2nd and 3rd selfed-generation). The sibling of a paternal *A. arenosa* parent (Aa4) and the maternal tetraploid *A. thaliana* parent (6978 aka "Wa-1") of our artificial *A. suecica* line were included as part of our samples (Supplementary Data 1). Each accession was replicated 3 times. Seeds were stratified in the dark for 4 days at 4°C in 1 ml of sterilised water. Seeds were then transferred to pots in a controlled growth chamber at 21°C. Humidity was kept constant at 60%. Pots were thinned to 2-3 seedlings after 1 week. Pots were re-randomized each week in their trays. Whole rosettes were collected when plants reached the 7-9 true-leaf stage of development. Samples were collected between 14:00-17:00h and flash-frozen in liquid nitrogen.

RNA extraction and library preparation

For each accession, 2-3 whole rosettes in each pot were pooled and total RNA was extracted using the ZR Plant RNA MiniPrepTM kit. We treated the samples with DNAse, and performed purification of mRNA and polyA selection using the AMPure XP magnetic beads and the Poly(A) RNA Selection Kit from Lexogen. RNA quality and degradation were assessed using the RNA Fragment Analyzer (DNF-471 stranded sensitivity RNA analysis kit, 15nt). Concentration of RNA per sample was measured using the Qubit fluorometer. Library preparation was carried out following the NEBNext Ultra II RNA Library Prep Kit for Illumina. Barcoded adaptors were ligated using NEBNext Multiplex Oligos for Illumina (Index Primers Set 1 and 2). The libraries were PCR amplified for 7 cycles. 125bp paired-end sequencing was carried out at the VBCF on Illumina (HiSeq 2500) using multiplexing.

RNA-seq mapping and gene expression analysis

We mapped 125bp paired-end reads to the *de novo* assembled *A. suecica* reference using STAR[154] (version 2.7), we filtered for primary and uniquely aligned reads using the parameters --outfilterMultimapNmax 1 --outSamprimaryFlag OneBestScore. We quantified reads mapped to genes using --quantMode GeneCounts.

In order to reduce signals that are the result of cross mapping between the subgenomes of *A. suecica* we used *A. thaliana* and *A. arenosa* as a control. For each gene in the *A. thaliana* subgenome we compared log fold change of gene counts in our *A. thaliana* population to those in our *A. arenosa* population. We filtered for genes with a $\log_2$(*A. thaliana*/*A. arenosa*) below 0. We applied the same filters for genes on the *A. arenosa* subgenome, here a $\log_2$(*A. arenosa*/*A. thaliana*) below 0. This reduced the number of genes analyzed from 22,383 to 21,737 on the *A. thaliana* subgenome, and 23,353 to 23,221 on the *A. arenosa* subgenome

Expression analysis was then further restricted to 1:1 unique homeologous gene pairs between the subgenomes of *A. suecica* (17,881 gene pairs). Gene counts were normalized for gene size by calculating Transcripts Per Million (TPM). The effective library sizes were calculated by computing a scaling factor based on the trimmed mean of M-values (TMM) in edgeR[155], separately for each subgenome. Lowly expressed genes were removed from the analysis by keeping genes that were expressed in at least 3 individuals of *A. thaliana* and *A. suecica*, at least 1 individual of *A. arenosa* and at least 1 individual of synthetic *A. suecica*. 14,041 homeologous gene pairs satisfied our expression criteria. Since *A. suecica* is expressing both subgenomes, in order to correctly normalize the effective library size in *A. suecica* accessions, the effective library size was calculated as a mean of TPM counts for both subgenomes. The effective library size of *A. thaliana* accessions was calculated for TPM counts using the *A. thaliana* subgenome of the reference genome, as genes from this subgenome will be expressed in *A. thaliana*, and the effective library size of *A. arenosa* lines using the *A. arenosa* subgenome of the reference *A. suecica* genome. Gene counts were transformed to count per million (CPM) with a prior count of 1, and were $\log_2$-transformed. We used the mean of replicates per accession for downstream analyses.

To compare homeologous genes between the subgenomes in *A. suecica* we computed a log-fold change using $\log_2$(*A. arenosa* homeolog/*A. thaliana* homeolog). For tissue-specific genes we took genes that showed a log-fold change >=2 in expression between two tissues.

For comparing homologous genes between the (sub-)genomes of *A. suecica* and the ancestral species *A. thaliana* and *A. arenosa*, we performed a Wilcoxon test independently for each of the 14,041 homeologous gene-pairs. Using the normalised CPM values, we compared the relative expression level of a gene on the *A. thaliana* subgenome between our population of *A. thaliana* and *A. suecica*. We performed the same test on the *A. arenosa* subgenome comparing relative expression of a gene between our population of *A. arenosa* and *A. suecica*. We filtered for genes with an adjusted p-value below <0.05 (using FDR correction). This amounted to 4,186 and 4,571 DEGs for the *A. thaliana* and *A. arenosa* subgenomes, respectively.

Cross-mapping of short reads

Cross-mapping of short RNA reads between the subgenomes of *A. suecia* was measured by mixing the RNA reads between *A. thaliana* and *A. arenosa* individuals to generate “in silico” *A. suecica* individuals. We mapped reads from 10 in-silico *A. suecica* individuals to the *A. suecica* genome. We compared different RNAseq pipelines to determine cross-mapping error rates. We mapped reads using STAR[154] (version 2.7), HISAT2[156] (version 2.1) and EAGLE[157]. ~1% of A. thaliana reads map to the A. arenosa subgenome and ~6% of the *A. arenosa* reads map to the *A. thaliana* subgenome, regardless of mapping strategy or pipeline (see Supplementary Figure 13). This can be explained by pairwise percentage differences or π within *A. arenosa* overlapping this distribution of π between *A. thaliana* and *A. arenosa* such that some exons on the *A. thaliana* subgenome are in fact closer to a particular *A. arenosa* individual than those on the *A. arenosa* subgenome of *A. suecica*. However lower π in *A. suecica* suggests this observation will not affect estimates of subgenome dominance for *A. suecica*.

Expression analysis of rRNA

RNA reads were mapped in a similar manner as DNA reads for the analysis of rDNA copy number (above). Expression analysis was performed in a similar manner to protein coding genes, in edgeR. We defined the exclusive expression of a particular 45S rRNA gene by taking a cut-off of 15 for $\log_2$(CPM) as this was the maximum level of cross-mapping we observed for the ancestral species (see Supplementary Fig. 6).

Expression analysis of transposable elements

To analyse the expression of transposable elements between species, the annotated TE consensus sequences in *A. suecica* were aligned using BLAST all vs all. Highly similar TE sequences (more than 85% similar for more than 85% percent of the TE sequence length), were removed, leaving 813 TE families out of 1213. Filtered *A. suecica* TEs were aligned to annotated *A. thaliana* (TAIR10) and *A. arenosa* (the PacBio contig assembly presented in this study) TE sequences to assign each family to an ancestral species using BLAST. 208 TE families were assigned to the *A. thaliana* parent and 171 TE families were assigned to the *A. arenosa* parent.

RNA reads were mapped to TE sequences using a similar approach as for gene expression analysis using edgeR. TEs that showed expression using a cut-off of $\log_2$CPM > 2 were kept. 121 *A. thaliana* TE sequences and 93 *A. arenosa* TE sequences passed this threshold. We took the mean of replicates per accession for further downstream analyses.

Gene ontology (GO) enrichment analysis

We used the R package TopGO[158] to conduct gene ontology enrichment analysis. We used the "weight01" algorithm when running TopGO which accounts for the hierarchical structure of GO terms and thus implicitly corrects for multiple testing. GO annotations were based on the *A. thaliana* ortholog of *A. suecica* genes. Gene annotations for *A. thaliana* were obtained using the R package biomaRt[159] from Ensembl 'biomaRt::useMart(biomart = "plants_mart", dataset = "athaliana_eg_gene", host = 'plants.ensembl.org').

Genome sizes measurements

We measured genome size for the reference *A. suecica* accession "ASS3" and the *A. arenosa* accession used for PacBio "Aa4", using *Solanum lycopersicum* cv. Stupicke (2C = 1.96 pg DNA) as the standard. The reference *A. lyrata* accession "MN47" and the *A. thaliana* accession "CVI" were used as additional controls. Each sample had 2 replicates.

In brief, the leaves from three week old fresh tissue were chopped using a razor blade in 500 µl of UV Precise P extraction buffer + 10 µl mercaptoethanol per ml (kit PARTEC CyStain PI Absolute P no. 05- 5022) to isolate nuclei. Instead of the Partec UV Precise P staining buffer, however, 1 ml of a 5 mg DAPI solution was used, as DAPI provides DNA content histograms with high resolution. The suspension was then passed through a 30 µm filter (Partec CellTrics no. 04-0042-2316) and incubated for 15 minutes on ice before FACs.

Genome size was measured using flow cytometry and a FACS Aria III sorter with near UV 375nm laser for DAPI. Debris was excluded by selecting peaks when plotting DAPI-W against DAPI-A for 20,000 events.

The data were analyzed using the flowCore[160] package in R. Genome size was estimated by comparing the mean G1 of the standard *Solanum lycopersicum* to that of each sample to calculate the 2C DNA content of that sample using the equation:

$$Sample\ 2C\ DNA\ content\ =\ [(sample\ G1\ peak\ mean)/(standard\ G1\ peak\ mean)] \\ *\ standard\ 2C\ DNA\ content$$

We also measured genome size for the reference *A. suecica* accession “ASS3” using the software jellyfish[161] and findGSE[162] using kmers (21mers). The genome size estimated was 312Mb, compared to the 305Mb estimated using FACs (see Supplementary Fig 1).

Mapping of TE insertions

We used PopoolationTE2[100] (version v1.10.04) to identify TE insertions. The advantage of this TE-calling software to others is that it avoids a reference bias by treating all TEs as *de-novo* insertions. Briefly, it works by using discordant read pairs to calculate the location and abundance of a TE in the genome for an accession of interest.

We mapped 100 bp Illumina DNA reads from [20,76,163], in addition to our newly generated synthetic *A. suecica* using BWA MEM[135] (version 0.7.15) to a repeat-masked version of the *A. suecica* reference genome, concatenated with our annotated repeat sequences (see ‘Genome annotation’), as this is the data format required by PopoolationTE2. Reads were given an increased penalty of 15 for being unpaired. Reads were de-duplicated using Samtools[136] rmdup (version 1.9). The resulting bam files were then provided to PopoolationTE2 to identify TE insertions in the genome of each of our *A. suecica*, *A. thaliana* and *A. arenosa* accessions. We used a mapping quality of 10 for the read in the discordant read pair mapping to the genome. We used the ‘separate’ mode in the ‘identify TE signatures’ step and a ‘--min-distance -200 --max-distance 500’ in the ‘pairupsignatues’ step of the pipeline. TE counts within each accession were merged if they fell within 400 bp of each other and if they mapped to the same TE sequence. All TE counts (i.e. the processed TE counts for each accession) were then combined to produce a population-wide count estimate. Population wide TE insertions were merged if they mapped to the same TE sequence and fell within 400 bp of each other. Coverage of each TE insertion in the population was also calculated for each accession. The final file was a list TE insertions present in the population and the presence or absence (or “NA” if there was no coverage to support the presence or absence of a TE insertion) in each accession analyzed (Supplementary Data 1).

Assigning ancestry to TE sequences

In order to examine TE consensus sequences that have mobilized between the subgenomes of *A. suecica*, we first examined which of our TE consensus sequences (N=1152) have at least the potential to mobilize (i.e. have full length TE copies in the genome of *A. suecica)*. We filtered for TE consensus sequences that had TE copies in the genome of *A. suecica* that are more than 80% similar in identity for more than 80% of the consensus sequence length (N=936). Of these, 188 consensus sequences were private to the *A. thaliana* subgenome, 460 were private to the *A. arenosa* subgenome, and 288 TE consensus sequences were present in both subgenomes of *A. suecica*. To determine if TEs have jumped from the *A. thaliana* subgenome to the *A. arenosa* subgenome and vice versa we next needed to assign ancestry to these 288 TE consensus sequences. To do this we used BLAST to search for these consensus sequences in the ancestral genomes of *A. suecica*, using the TAIR10 *A. thaliana* reference and our *A. arenosa* PacBio contig assembly. Using the same 80%-80% rule we assigned 55 TEs to *A. arenosa* and 15 TEs to *A. thaliana* ancestry.

Read mapping and SNP calling

To call biallelic SNPs we mapped reads to the *A. suecica* reference genome using the same filtering parameters described in “Mapping of TE insertions”. Biallelic SNPs were called using HaplotypeCaller from GATK[164] (version 3.8) using default quality thresholds. SNPs were annotated using SnpEff[165]. Biallelic SNPs on the *A. thaliana* sub-genome were polarized using 38 diploid *A. lyrata* lines[76] and biallelic SNPS on the *A. arenosa* sub-genome were polarized using 30 *A. thaliana* accessions[163] closely related to *A. suecica*[20].

Chromosome preparation and FISH

Whole inflorescences of *A. arenosa*, *A. suecica* and *A. thaliana* were fixed in freshly prepared ethanol:acetic acid fixative (3:1) overnight, transferred into 70% ethanol and stored at -20°C until use. Selected inflorescences were rinsed in distilled water and citrate buffer (10 mM sodium citrate, pH 4.8), and digested by a 0.3% mix of pectolytic enzymes (cellulase, cytohelicase, pectolyase; all from Sigma-Aldrich) in citrate buffer for c. 3 hrs. Mitotic chromosome spreads were prepared from pistils as previously described[166] by Mandáková and Lysak and suitable slides pretreated by RNase (100 µg/ml, AppliChem) and pepsin (0.1 mg/ml, Sigma-Aldrich).

For identification of *A. thaliana* and *A. arenosa* subgenomes in the allotetraploid genome of *A. suecica*, FISH probes were made from plasmids pARR20–1 or pAaCEN containing 180 bp of *A. thaliana* (pAL; Vongs et al. 1993) or ~250 bp of *A. arenosa* (pAa; Kamm et al. 1995) pericentromeric repeats, respectively. The *A. thaliana* BAC clone T15P10 (AF167571) bearing 45S rRNA gene repeats was used for in situ localization of NORs. Individual probes were labeled with biotin-dUTP, digoxigenin-dUTP and Cy3-dUTP by nick translation, pooled, precipitated, and resuspended in 20 µl of hybridization mixture [50% formamide and 10% dextran sulfate in 2× saline sodium citrate (2× SSC)] per slide as previously described[96].

Probes and chromosomes were denatured together on a hot plate at 80°C for 2 min and incubated in a moist chamber at 37°C overnight. Post hybridization washing was performed in 20% formamide in 2× SSC at 42°C. Fluorescent detection was as follows: biotin-dUTP was detected by avidin–Texas Red (Vector Laboratories) and amplified by goat anti-avidin–biotin (Vector Laboratories) and avidin–Texas Red; digoxigenin-dUTP was detected by mouse anti-digoxigenin (Jackson ImmunoResearch) and goat anti-mouse Alexa Fluor 488 (Molecular Probes). Chromosomes were counterstained with DAPI (4’,6-diamidino-2-phenylindole; 2 µg/ml) in Vectashield (Vector Laboratories). Fluorescent signals were analyzed and photographed using a Zeiss Axioimager epifluorescence microscope and a CoolCube camera (MetaSystems). Images were acquired separately for the four fluorochromes using appropriate excitation and emission filters (AHF Analysentechnik). The monochromatic images were pseudo colored and merged using Adobe Photoshop CS6 software (Adobe Systems).

DAP-seq enrichment analysis for transcription factor target genes

We downloaded the target genes of transcription factors from the plant cistrome database (http://neomorph.salk.edu/dap_web/pages/index.php), which is a collection of transcription factor binding sites and their target genes, in *A. thaliana*, based on DAP-seq[167]. To test for enrichment of a gene set (for example the genes in *A. thaliana* cluster 2 on Fig. 5) for target genes of a particular transcription factor, we performed a hyper-geometric test in R. As a background we used the total 14,041 genes used in our gene expression analysis. We then performed FDR correction for multiple testing to calculate an accurate p-value of the enrichment.

Data Availability

Genome assemblies and raw short reads can be found in the European Nucleotide Archive (ENA) (https://www.ebi.ac.uk/ena/browser/home).
The genome assembly for *A. suecica* ASS3 can be found under the BioProject number PRJEB42198, assembly accession GCA_905175345. The raw reads for the *A. suecica* genome assembly generated by Pacbio RSII can be found under ERR5037702 and those from Sequel under ERR5031296. The HiC reads used for scaffolding the *A. suecica* assembly can be found under ERR5032369.
The contig assembly for tetraploid *A. arenosa (ssp. arenosa)* can be found under the BioProject number PRJEB42276, assembly accession GCA_905175405. The raw reads for the *A. arenosa* Aa4 contig assembly generated by Sequel can be found under ERR5031542 and the reads generated by Nanopore under ERR5031541. HiC reads for the *A. arenosa* assembly can be found under ERR5032370.
HiC sequencing data for the ancestral species, the outlier accession AS530 and synthetic *A. suecica* can be found under the BioProject PRJEB42290.
DNA resequencing of synthetic *A. suecica* and parents generated in this study can be found under the BioProject PRJEB42291.
The RNA-seq reads are under the BioProject number PRJEB42277.
TE presence/absence calls for *A. suecica* and the ancestral species can be found in Supplementary Data 1.
A list of DEGs, orthologs, enriched DAP-seq transcription factors, CyMIRA gene overlaps and RNA-seq mapping statistics can be found in Supplementary Data 2.
Log fold change and CPM (counts per million) for genes on the *A. thaliana* and *A. arenosa* subgenome can be found in Supplementary Data 3.
The gene annotation (gff3 file) of the *A. suecica* genome can be found in Supplementary Data 4.
TE consensus sequences and a hierarchy file of TE order for *A. suecica* can be found in Supplementary Data 5.

Acknowledgments

This work was supported, in part, by DFG SPP 1529 to M.N. and Detlef Weigel. T.M. and M.A.L. were supported by the Czech Science Foundation (grant no. 19-03442S) and the CEITEC 2020 project (grant no. LQ1601). P.Y.N. acknowledges postdoctoral fellowship of the

Research Foundation–Flanders (12S9618N). We thank the Next Generation Sequencing Unit of the Vienna Biocenter Core Facilities (VBCF) for assistance. We thank Svante Holm and Torbjörn Säll for material collections and helpful discussions throughout. We also thank Yves Van de Peer for providing useful feedback on the manuscript, and Joel Sharbrough for pointing us to the CyMIRA database

References

1 Van de Peer Y, Mizrachi E, Marchal K. The evolutionary significance of polyploidy. *Nat Rev Genet* 2017; **18**: 411–424.

2 Soltis PS, Soltis DE. Ancient WGD events as drivers of key innovations in angiosperms. *Curr Opin Plant Biol* 2016; **30**: 159–165.

3 Dehal P, Boore JL. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol* 2005; **3**: e314.

4 Li Z, Tiley GP, Galuska SR, Reardon CR, Kidder TI, Rundell RJ *et al.* Multiple large-scale gene and genome duplications during the evolution of hexapods. *Proc Natl Acad Sci U S A* 2018; **115**: 4713–4718.

5 Chen ZJ, Sreedasyam A, Ando A, Song Q, De Santiago LM, Hulse-Kemp AM *et al.* Genomic diversifications of five Gossypium allopolyploid species and their impact on cotton improvement. *Nat Genet* 2020; **52**: 525–533.

6 Edger PP, Poorten TJ, VanBuren R, Hardigan MA, Colle M, McKain MR *et al.* Origin and evolution of the octoploid strawberry genome. *Nat Genet* 2019; **51**: 541–547.

7 Ramírez-González RH, Borrill P, Lang D, Harrington SA, Brinton J, Venturini L *et al.* The transcriptional landscape of polyploid wheat. *Science* 2018; **361**. doi:10.1126/science.aar6089.

8 Zhuang W, Chen H, Yang M, Wang J, Pandey MK, Zhang C *et al.* The genome of cultivated peanut provides insight into legume karyotypes, polyploid evolution and crop domestication. Nature Genetics. 2019; **51**: 865–876.

9 Bertioli DJ, Jenkins J, Clevenger J, Dudchenko O, Gao D, Seijo G *et al.* The genome sequence of segmental allotetraploid peanut Arachis hypogaea. *Nat Genet* 2019; **51**: 877–884.

10 Kasianov AS, Klepikova AV, Kulakovskiy IV, Gerasimov ES, Fedotova AV, Besedina EG *et al.* High-quality genome assembly of Capsella bursa-pastoris reveals asymmetry of regulatory elements at early stages of polyploid genome evolution. *Plant J* 2017; **91**: 278–291.

11 Kryvokhyzha D, Milesi P, Duan T, Orsucci M, Wright SI, Glémin S *et al.* Towards the new normal: Transcriptomic convergence and genomic legacy of the two subgenomes of an allopolyploid weed (Capsella bursa-pastoris). *PLoS Genet* 2019; **15**: e1008131.

12 Douglas GM, Gos G, Steige KA, Salcedo A, Holm K, Josephs EB *et al.* Hybrid origins and the earliest stages of diploidization in the highly successful recent polyploid Capsella bursa-pastoris. *Proc Natl Acad Sci U S A* 2015; **112**: 2806–2811.

13 Griffiths AG, Moraga R, Tausen M, Gupta V, Bilton TP, Campbell MA *et al.* Breaking Free: The Genomics of Allopolyploidy-Facilitated Niche Expansion in White Clover. *Plant Cell* 2019; **31**: 1466–1487.

14 Gordon SP, Contreras-Moreira B, Levy JJ, Djamei A, Czedik-Eysenberg A, Tartaglio VS *et al.* Gradual polyploid genome evolution revealed by pan-genomic analysis of Brachypodium

hybridum and its diploid progenitors. *Nat Commun* 2020; **11**: 3670.

15 Catalán P, López-Álvarez D, Bellosta C, Villar L. Updated taxonomic descriptions, iconography, and habitat preferences of Brachypodium distachyon, B. stacei , and B. hybridum (Poaceae). *An Jard Bot Madr* 2016; **73**: 028.

16 Paape T, Briskine RV, Halstead-Nussloch G, Lischer HEL, Shimizu-Inatsugi R, Hatakeyama M *et al.* Patterns of polymorphism and selection in the subgenomes of the allopolyploid Arabidopsis kamchatica. *Nat Commun* 2018; **9**: 3909.

17 Edger PP, Smith R, McKain MR, Cooley AM, Vallejo-Marin M, Yuan Y *et al.* Subgenome Dominance in an Interspecific Hybrid, Synthetic Allopolyploid, and a 140-Year-Old Naturally Established Neo-Allopolyploid Monkeyflower. *Plant Cell* 2017; **29**: 2150–2167.

18 Soltis DE, Soltis PS, Pires JC, Kovarik A, Tate JA, Mavrodiev E. Recent and recurrent polyploidy in Tragopogon (Asteraceae): cytogenetic, genomic and genetic comparisons. *Biol J Linn Soc Lond* 2004; **82**: 485–501.

19 te Beest M, Le Roux JJ, Richardson DM, Brysting AK, Suda J, Kubesová M *et al.* The more the better? The role of polyploidy in facilitating plant invasions. *Ann Bot* 2012; **109**: 19–45.

20 Novikova PY, Tsuchimatsu T, Simon S, Nizhynska V, Voronin V, Burns R *et al.* Genome Sequencing Reveals the Origin of the Allotetraploid Arabidopsis suecica. *Mol Biol Evol* 2017; **34**: 957–968.

21 Fowler NL, Levin DA. Ecological Constraints on the Establishment of a Novel Polyploid in Competition with Its Diploid Progenitor. *Am Nat* 1984; **124**: 703–711.

22 Bomblies K, Madlung A. Polyploidy in the Arabidopsis genus. *Chromosome Res* 2014; **22**: 117–134.

23 Hollister JD, Arnold BJ, Svedin E, Xue KS, Dilkes BP, Bomblies K. Genetic adaptation associated with genome-doubling in autotetraploid Arabidopsis arenosa. *PLoS Genet* 2012; **8**: e1003093.

24 Bomblies K, Jones G, Franklin C, Zickler D, Kleckner N. The challenge of evolving stable polyploidy: could an increase in 'crossover interference distance' play a central role? *Chromosoma* 2016; **125**: 287–300.

25 Leitch AR, Leitch IJ. Genomic plasticity and the diversity of polyploid plants. *Science* 2008; **320**: 481–483.

26 Bottani S, Zabet NR, Wendel JF, Veitia RA. Gene Expression Dominance in Allopolyploids: Hypotheses and Models. *Trends Plant Sci* 2018; **23**: 393–402.

27 Parisod C, Alix K, Just J, Petit M, Sarilar V, Mhiri C *et al.* Impact of transposable elements on the organization and function of allopolyploid genomes. *New Phytol* 2010; **186**: 37–45.

28 McClintock B. The significance of responses of the genome to challenge. Science. 1984; **226**: 792–801.

29 Feldman M, Liu B, Segal G, Abbo S, Levy AA, Vega JM. Rapid elimination of low-copy DNA sequences in polyploid wheat: a possible mechanism for differentiation of homoeologous chromosomes. *Genetics* 1997; **147**: 1381–1387.

30 Zhang H, Gou X, Zhang A, Wang X, Zhao N, Dong Y *et al.* Transcriptome shock invokes disruption of parental expression-conserved genes in tetraploid wheat. *Sci Rep* 2016; **6**: 26363.

31 Wang X, Zhang H, Li Y, Zhang Z, Li L, Liu B. Transcriptome asymmetry in synthetic and natural

allotetraploid wheats, revealed by RNA-sequencing. *New Phytol* 2016; **209**: 1264–1277.

32 Zhang H, Bian Y, Gou X, Zhu B, Xu C, Qi B *et al.* Persistent whole-chromosome aneuploidy is generally associated with nascent allohexaploid wheat. *Proc Natl Acad Sci U S A* 2013; **110**: 3447–3452.

33 Kashkush K, Feldman M, Levy AA. Gene loss, silencing and activation in a newly synthesized wheat allotetraploid. *Genetics* 2002; **160**: 1651–1659.

34 Shaked H, Kashkush K, Ozkan H, Feldman M, Levy AA. Sequence elimination and cytosine methylation are rapid and reproducible responses of the genome to wide hybridization and allopolyploidy in wheat. *Plant Cell* 2001; **13**: 1749–1759.

35 Ozkan H, Levy AA, Feldman M. Allopolyploidy-Induced Rapid Genome Evolution in the Wheat (Aegilops–Triticum) Group. *Plant Cell* 2001; **13**: 1735–1747.

36 Xiong Z, Gaeta RT, Pires JC. Homoeologous shuffling and chromosome compensation maintain genome balance in resynthesized allopolyploid Brassica napus. *Proc Natl Acad Sci U S A* 2011; **108**: 7908–7913.

37 Wu J, Lin L, Xu M, Chen P, Liu D, Sun Q *et al.* Homoeolog expression bias and expression level dominance in resynthesized allopolyploid Brassica napus. *BMC Genomics* 2018; **19**: 586.

38 Szadkowski E, Eber F, Huteau V, Lodé M, Huneau C, Belcram H *et al.* The first meiosis of resynthesized Brassica napus, a genome blender. *New Phytol* 2010; **186**: 102–112.

39 Zhao T, Tao X, Feng S, Wang L, Hong H, Ma W *et al.* LncRNAs in polyploid cotton interspecific hybrids are derived from transposon neofunctionalization. *Genome Biol* 2018; **19**: 195.

40 Yoo M-J, Szadkowski E, Wendel JF. Homoeolog expression bias and expression level dominance in allopolyploid cotton. *Heredity* 2013; **110**: 171–180.

41 Li A, Liu D, Wu J, Zhao X, Hao M, Geng S *et al.* mRNA and Small RNA Transcriptomes Reveal Insights into Dynamic Homoeolog Regulation of Allopolyploid Heterosis in Nascent Hexaploid Wheat. *Plant Cell* 2014; **26**: 1878–1900.

42 Flagel LE, Wendel JF. Evolutionary rate variation, genomic dominance and duplicate gene expression evolution during allotetraploid cotton speciation. *New Phytol* 2010; **186**: 184–193.

43 Liu B, Brubaker CL, Mergeai G, Cronn RC, Wendel JF. Polyploid formation in cotton is not accompanied by rapid genomic changes. *Genome* 2001; **44**: 321–330.

44 Kashkush K, Feldman M, Levy AA. Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat. *Nat Genet* 2003; **33**: 102–106.

45 Kraitshtein Z, Yaakov B, Khasdan V, Kashkush K. Genetic and epigenetic dynamics of a retrotransposon after allopolyploidization of wheat. *Genetics* 2010; **186**: 801–812.

46 Yaakov B, Kashkush K. Mobilization of Stowaway-like MITEs in newly formed allohexaploid wheat species. *Plant Mol Biol* 2012; **80**: 419–427.

47 International Wheat Genome Sequencing Consortium (IWGSC), IWGSC RefSeq principal investigators:, Appels R, Eversole K, Feuillet C, Keller B *et al.* Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science* 2018; **361**. doi:10.1126/science.aar7191.

48 Wang M, Tu L, Yuan D, Zhu D, Shen C, Li J *et al.* Reference genome sequences of two cultivated allotetraploid cottons, Gossypium hirsutum and Gossypium barbadense. *Nat Genet* 2019; **51**: 224–229.

49 Yang Z, Ge X, Yang Z, Qin W, Sun G, Wang Z *et al.* Extensive intraspecific gene order and gene structural variations in upland cotton cultivars. *Nat Commun* 2019; **10**: 2989.

50 Huang G, Wu Z, Percy RG, Bai M, Li Y, Frelichowski JE *et al.* Genome sequence of Gossypium herbaceum and genome updates of Gossypium arboreum and Gossypium hirsutum provide insights into cotton A-genome evolution. *Nat Genet* 2020; **52**: 516–524.

51 Zhang T, Hu Y, Jiang W, Fang L, Guan X, Chen J *et al.* Sequencing of allotetraploid cotton (Gossypium hirsutum L. acc. TM-1) provides a resource for fiber improvement. *Nat Biotechnol* 2015; **33**: 531–537.

52 Han J, Masonbrink RE, Shan W, Song F, Zhang J, Yu W *et al.* Rapid proliferation and nucleolar organizer targeting centromeric retrotransposons in cotton. *Plant J* 2016; **88**: 992–1005.

53 Wang M, Wang P, Lin M, Ye Z, Li G, Tu L *et al.* Evolutionary dynamics of 3D genome architecture following polyploidization in cotton. *Nat Plants* 2018; **4**: 90–97.

54 Cheng F, Wu J, Fang L, Sun S, Liu B, Lin K *et al.* Biased gene fractionation and dominant gene expression among the subgenomes of Brassica rapa. *PLoS One* 2012; **7**: e36442.

55 Schnable JC, Springer NM, Freeling M. Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proc Natl Acad Sci U S A* 2011; **108**: 4069–4074.

56 International Wheat Genome Sequencing Consortium (IWGSC). A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. *Science* 2014; **345**: 1251788.

57 Chalhoub B, Denoeud F, Liu S, Parkin IAP, Tang H, Wang X *et al.* Plant genetics. Early allopolyploid evolution in the post-Neolithic Brassica napus oilseed genome. *Science* 2014; **345**: 950–953.

58 Wang M, Tu L, Lin M, Lin Z, Wang P, Yang Q *et al.* Asymmetric subgenome selection and cis-regulatory divergence during cotton domestication. *Nat Genet* 2017; **49**: 579–587.

59 Gaut BS, Seymour DK, Liu Q, Zhou Y. Demography and its effects on genomic variation in crop domestication. *Nat Plants* 2018; **4**: 512–520.

60 Kremling KAG, Chen S-Y, Su M-H, Lepak NK, Romay MC, Swarts KL *et al.* Dysregulation of expression correlates with rare-allele burden and fitness loss in maize. *Nature* 2018; **555**: 520–523.

61 Qian L, Qian W, Snowdon RJ. Sub-genomic selection patterns as a signature of breeding in the allopolyploid Brassica napus genome. *BMC Genomics* 2014; **15**: 1170.

62 Wang L, Beissinger TM, Lorant A, Ross-Ibarra C, Ross-Ibarra J, Hufford MB. The interplay of demography and selection during maize domestication and expansion. *Genome Biol* 2017; **18**: 215.

63 Alonge M, Wang X, Benoit M, Soyk S, Pereira L, Zhang L *et al.* Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato. *Cell* 2020; **182**: 145–161.e23.

64 Liu Y, Du H, Li P, Shen Y, Peng H, Liu S *et al.* Pan-Genome of Wild and Cultivated Soybeans. *Cell* 2020; **182**: 162–176.e13.

65 Zhou Y, Minio A, Massonnet M, Solares E, Lv Y, Beridze T *et al.* The population genetics of structural variants in grapevine domestication. *Nat Plants* 2019; **5**: 965–979.

66 Buggs RJA, Zhang L, Miles N, Tate JA, Gao L, Wei W *et al.* Transcriptomic shock generates evolutionary novelty in a newly formed, natural allopolyploid plant. *Curr Biol* 2011; **21**: 551–556.

67 Chester M, Gallagher JP, Symonds VV, Cruz da Silva AV, Mavrodiev EV, Leitch AR *et al.* Extensive chromosomal variation in a recently formed natural allopolyploid species, Tragopogon miscellus (Asteraceae). *Proc Natl Acad Sci U S A* 2012; **109**: 1176–1181.

68 Chelaifa H, Monnier A, Ainouche M. Transcriptomic changes following recent natural hybridization and allopolyploidy in the salt marsh species Spartina × townsendii and Spartina anglica (Poaceae). *New Phytol* 2010; **186**: 161–174.

69 Kryvokhyzha D, Salcedo A, Eriksson MC, Duan T, Tawari N, Chen J *et al.* Parental legacy, demography, and admixture influenced the evolution of the two subgenomes of the tetraploid Capsella bursa-pastoris (Brassicaceae). *PLoS Genet* 2019; **15**: e1007949.

70 Akama S, Shimizu-Inatsugi R, Shimizu KK, Sese J. Genome-wide quantification of homeolog expression ratio revealed nonstochastic gene regulation in synthetic allopolyploid Arabidopsis. *Nucleic Acids Res* 2014; **42**: e46.

71 Wu H, Yu Q, Ran J-H, Wang X-Q. Unbiased subgenome evolution in allotetraploid species of Ephedra and its implications for the evolution of large genomes in gymnosperms. *Genome Biol Evol* 2020. doi:10.1093/gbe/evaa236.

72 Säll T, Lind-Halldén C, Jakobsson M, Halldén C. Mode of reproduction in Arabidopsis suecica. *Hereditas* 2004; **141**: 313–317.

73 Hohmann N, Wolf EM, Lysak MA, Koch MA. A Time-Calibrated Road Map of Brassicaceae Species Radiation and Evolutionary History. *Plant Cell* 2015; **27**: 2770–2784.

74 O’Kane SL, Schaal BA, Al-Shehbaz IA. The Origins of Arabidopsis suecica (Brassicaceae) as Indicated by Nuclear rDNA Sequences. *Syst Bot* 1996; **21**: 559–566.

75 Jakobsson M, Hagenblad J, Tavaré S, Säll T, Halldén C, Lind-Halldén C *et al.* A unique recent origin of the allotetraploid species Arabidopsis suecica: Evidence from nuclear DNA markers. *Mol Biol Evol* 2006; **23**: 1217–1231.

76 Novikova PY, Hohmann N, Nizhynska V, Tsuchimatsu T, Ali J, Muir G *et al.* Sequencing of the genus Arabidopsis identifies a complex history of nonbifurcating speciation and abundant trans-specific polymorphism. *Nat Genet* 2016; **48**: 1077–1082.

77 Slotte T, Hazzouri KM, Ågren JA, Koenig D, Maumus F, Guo Y-L *et al.* The Capsella rubella genome and the genomic consequences of rapid mating system evolution. *Nat Genet* 2013; **45**: 831–835.

78 Liu S, Liu Y, Yang X, Tong C, Edwards D, Parkin IAP *et al.* The Brassica oleracea genome reveals the asymmetrical evolution of polyploid genomes. *Nat Commun* 2014; **5**: 3930.

79 Madlung A, Tyagi AP, Watson B, Jiang H, Kagochi T, Doerge RW *et al.* Genomic changes in synthetic Arabidopsis polyploids. *Plant J* 2005; **41**: 221–230.

80 Copenhaver GP, Pikaard CS. Two-dimensional RFLP analyses reveal megabase-sized clusters of rRNA gene variants in Arabidopsis thaliana, suggesting local spreading of variants as the mode for gene homogenization during concerted evolution. The Plant Journal. 1996; **9**: 273–282.

81 Navashin M. Chromosome Alterations Caused by Hybridization and Their Bearing upon Certain General Genetic Problems. *Cytologia* 1934; **5**: 169–203.

82 Tucker S, Vitins A, Pikaard CS. Nucleolar dominance and ribosomal RNA gene silencing. *Curr*

Opin Cell Biol 2010; **22**: 351–356.

83 Maciak S, Michalak K, Kale SD, Michalak P. Nucleolar Dominance and Repression of 45S Ribosomal RNA Genes in Hybrids between Xenopus borealis and X. muelleri (2n = 36). Cytogenetic and Genome Research. 2016; **149**: 290–296.

84 Książczyk T, Kovarik A, Eber F, Huteau V, Khaitova L, Tesarikova Z *et al.* Immediate unidirectional epigenetic reprogramming of NORs occurs independently of rDNA rearrangements in synthetic and natural forms of a polyploid species Brassica napus. Chromosoma. 2011; **120**: 557–571.

85 Chen ZJ, Comai L, Pikaard CS. Gene dosage and stochastic effects determine the severity and direction of uniparental ribosomal RNA gene silencing (nucleolar dominance) in Arabidopsis allopolyploids. *Proc Natl Acad Sci U S A* 1998; **95**: 14891–14896.

86 Pontes O, Lawrence RJ, Silva M, Preuss S, Costa-Nunes P, Earley K *et al.* Postembryonic establishment of megabase-scale gene silencing in nucleolar dominance. *PLoS One* 2007; **2**: e1157.

87 Lewis MS, Pikaard CS. Restricted chromosomal silencing in nucleolar dominance. *Proc Natl Acad Sci U S A* 2001; **98**: 14536–14540.

88 Pontes O, Neves N, Silva M, Lewis MS, Madlung A, Comai L *et al.* Chromosomal locus rearrangements are a rapid response to formation of the allotetraploid Arabidopsis suecica genome. Proceedings of the National Academy of Sciences. 2004; **101**: 18240–18245.

89 Long Q, Rabanal FA, Meng D, Huber CD, Farlow A, Platzer A *et al.* Massive genomic variation and strong selection in Arabidopsis thaliana lines from Sweden. *Nat Genet* 2013; **45**: 884–890.

90 Rabanal FA, Mandáková T, Soto-Jiménez LM, Greenhalgh R, Parrott DL, Lutzmayer S *et al.* Epistatic and allelic interactions control expression of ribosomal RNA gene clusters in Arabidopsis thaliana. *Genome Biol* 2017; **18**: 75.

91 Pontes O, Lawrence RJ, Neves N, Silva M, Lee J-H, Chen ZJ *et al.* Natural variation in nucleolar dominance reveals the relationship between nucleolus organizer chromatin topology and rRNA gene transcription in Arabidopsis. *Proc Natl Acad Sci U S A* 2003; **100**: 11418–11423.

92 Guo X, Han F. Asymmetric epigenetic modification and elimination of rDNA sequences by polyploidization in wheat. *Plant Cell* 2014; **26**: 4311–4327.

93 Liu B, Davis TM. Conservation and loss of ribosomal RNA gene sites in diploid and polyploid Fragaria (Rosaceae). *BMC Plant Biol* 2011; **11**: 1–13.

94 Steige KA, Slotte T. Genomic legacies of the progenitors and the evolutionary consequences of allopolyploidy. *Curr Opin Plant Biol* 2016; **30**: 88–93.

95 Vicient CM, Casacuberta JM. Impact of transposable elements on polyploid plant genomes. *Ann Bot* 2017; **120**: 195–207.

96 Ungerer MC, Strakosh SC, Zhen Y. Genome expansion in three hybrid sunflower species is associated with retrotransposon proliferation. *Curr Biol* 2006; **16**: R872–3.

97 Rieseberg LH, Raymond O, Rosenthal DM, Lai Z, Livingstone K, Nakazato T *et al.* Major ecological transitions in wild sunflowers facilitated by hybridization. *Science* 2003; **301**: 1211–1216.

98 Cavrak VV, Lettner N, Jamge S, Kosarewicz A, Bayer LM, Mittelsten Scheid O. How a retrotransposon exploits the plant's heat stress response for its activation. *PLoS Genet* 2014; **10**: e1004115.

99 Göbel U, Arce AL, He F, Rico A, Schmitz G, de Meaux J. Robustness of Transposable Element Regulation but No Genomic Shock Observed in Interspecific Arabidopsis Hybrids. *Genome Biol Evol* 2018; **10**: 1403–1415.

100 Kofler R, Gomez-Sanchez D, Schlotterer C. PoPoolationTE2: Comparative Population Genomics of Transposable Elements Using Pool-Seq. *Mol Biol Evol* 2016; **33**: 2759–2764.

101 Lockton S, Gaut BS. The evolution of transposable elements in natural populations of self-fertilizing Arabidopsis thaliana and its outcrossing relative Arabidopsis lyrata. *BMC Evol Biol* 2010; **10**: 10.

102 Quadrana L, Bortolini Silveira A, Mayhew GF, LeBlanc C, Martienssen RA, Jeddeloh JA *et al.* The Arabidopsis thaliana mobilome and its impact at the species level. *Elife* 2016; **5**. doi:10.7554/eLife.15716.

103 Stuart T, Eichten SR, Cahn J, Karpievitch YV, Borevitz JO, Lister R. Population scale mapping of transposable element diversity reveals links to gene regulation and epigenomic variation. *Elife* 2016; **5**. doi:10.7554/eLife.20777.

104 Wolfe KH. Yesterday's polyploids and the mystery of diploidization. *Nat Rev Genet* 2001; **2**: 333–341.

105 Conant GC, Birchler JA, Pires JC. Dosage, duplication, and diploidization: clarifying the interplay of multiple models for duplicate gene evolution over time. *Curr Opin Plant Biol* 2014; **19**: 91–98.

106 Aköz G, Nordborg M. The Aquilegia genome reveals a hybrid origin of core eudicots. *Genome Biol* 2019; **20**: 256.

107 Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE *et al.* Ancestral polyploidy in seed plants and angiosperms. *Nature* 2011; **473**: 97–100.

108 Soltis PS, Marchant DB, Van de Peer Y, Soltis DE. Polyploidy and genome evolution in plants. *Curr Opin Genet Dev* 2015; **35**: 119–125.

109 Thomas BC, Pedersen B, Freeling M. Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. *Genome Res* 2006; **16**: 934–946.

110 Renny-Byfield S, Gong L, Gallagher JP, Wendel JF. Persistence of subgenomes in paleopolyploid cotton after 60 my of evolution. *Mol Biol Evol* 2015; **32**: 1063–1071.

111 Garsmeur O, Schnable JC, Almeida A, Jourda C, D'Hont A, Freeling M. Two evolutionarily distinct classes of paleopolyploidy. *Mol Biol Evol* 2014; **31**: 448–454.

112 Li Q, Qiao X, Yin H, Zhou Y, Dong H, Qi K *et al.* Unbiased subgenome evolution following a recent whole-genome duplication in pear (Pyrus bretschneideri Rehd.). *Hortic Res* 2019; **6**: 34.

113 Shan S, Boatwright JL, Liu X, Chanderbali AS, Fu C, Soltis PS *et al.* Transcriptome Dynamics of the Inflorescence in Reciprocally Formed Allopolyploid Tragopogon miscellus (Asteraceae). *Front Genet* 2020; **11**: 888.

114 Bird KA, Niederhuth C, Ou S, Gehan M, Chris Pires J, Xiong Z *et al.* Replaying the evolutionary tape to investigate subgenome dominance in allopolyploid Brassica napus. doi:10.1101/814491.

115 Alger EI, Edger PP. One subgenome to rule them all: underlying mechanisms of subgenome dominance. *Curr Opin Plant Biol* 2020; **54**: 108–113.

116 Carlson KD, Fernandez-Pozo N, Bombarely A, Pisupati R, Mueller LA, Madlung A. Natural variation in stress response gene activity in the allopolyploid Arabidopsis suecica. *BMC*

Genomics 2017; **18**: 653.

117 Chang PL, Dilkes BP, McMahon M, Comai L, Nuzhdin SV. Homoeolog-specific retention and use in allotetraploid Arabidopsis suecica depends on parent of origin and network partners. *Genome Biol* 2010; **11**: R125.

118 Adams KL, Percifield R, Wendel JF. Organ-specific silencing of duplicated genes in a newly synthesized cotton allotetraploid. *Genetics* 2004; **168**: 2217–2226.

119 Sicard A, Lenhard M. The selfing syndrome: a model for studying the genetic and evolutionary basis of morphological adaptation in plants. *Ann Bot* 2011; **107**: 1433–1443.

120 Lu Y-J, Swamy KBS, Leu J-Y. Experimental Evolution Reveals Interplay between Sch9 and Polyploid Stability in Yeast. *PLoS Genet* 2016; **12**: e1006409.

121 Yant L, Hollister JD, Wright KM, Arnold BJ, Higgins JD, Franklin FC *et al.* Meiotic adaptation to genome duplication in Arabidopsis arenosa. *Curr Biol* 2013; **23**: 2151–2156.

122 Morgan C, Zhang H, Henry CE, Franklin FCH, Bomblies K. Derived alleles of two axis proteins affect meiotic traits in autotetraploid Arabidopsis arenosa. *Proc Natl Acad Sci U S A* 2020; **117**: 8980–8988.

123 Haga N, Kobayashi K, Suzuki T, Maeo K, Kubo M, Ohtani M *et al.* Mutations in MYB3R1 and MYB3R4 cause pleiotropic developmental defects and preferential down-regulation of multiple G2/M-specific genes in Arabidopsis. *Plant Physiol* 2011; **157**: 706–717.

124 Forsythe ES, Sharbrough J, Havird JC, Warren JM, Sloan DB. CyMIRA: The Cytonuclear Molecular Interactions Reference for Arabidopsis. *Genome Biol Evol* 2019; **11**: 2194–2202.

125 Wu Y, Lin F, Zhou Y, Wang J, Sun S, Wang B *et al.* Genomic mosaicism due to homoeologous exchange generates extensive phenotypic diversity in nascent allopolyploids. *Natl Sci Rev* 2020. doi:10.1093/nsr/nwaa277.

126 Darwin C. The origin of species by means of natural selection : or The preservation of favored races in the struggle for life / by Charles Darwin. 1872. doi:10.5962/bhl.title.2106.

127 Chin C-S, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A *et al.* Phased diploid genome assembly with single-molecule real-time sequencing. *Nat Methods* 2016; **13**: 1050–1054.

128 Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res* 2017; **27**: 722–736.

129 Chakraborty M, Baldwin-Brown JG, Long AD, Emerson JJ. Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. *Nucleic Acids Res* 2016; **44**: e147.

130 Chin C-S, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nature Methods. 2013; **10**: 563–569.

131 Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 2014; **9**: e112963.

132 Wingett S, Ewels P, Furlan-Magaril M, Nagano T, Schoenfelder S, Fraser P *et al.* HiCUP: pipeline for mapping and processing Hi-C data. *F1000Res* 2015; **4**: 1310.

133 Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. MUMmer4: A fast and versatile genome alignment system. *PLoS Comput Biol* 2018; **14**: e1005944.

134 Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, Shendure J. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat Biotechnol* 2013; **31**: 1119–1125.

135 Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009; **25**: 1754–1760.

136 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009; **25**: 2078–2079.

137 Himmelmann L. HMM: Hidden Markov Models. *R package version* 2010; **1**.

138 Broman KW, Wu H, Sen S, Churchill GA. R/qtl: QTL mapping in experimental crosses. *Bioinformatics* 2003; **19**: 889–890.

139 Durand NC, Shamim MS, Machol I, Rao SSP, Huntley MH, Lander ES *et al.* Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst* 2016; **3**: 95–98.

140 Stanke M, Morgenstern B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res* 2005; **33**: W465–7.

141 Seppey M, Manni M, Zdobnov EM. BUSCO: Assessing Genome Assembly and Annotation Completeness. In: Kollmar M (ed). *Gene Prediction: Methods and Protocols*. Springer New York: New York, NY, 2019, pp 227–245.

142 Rawat V, Abdelsamad A, Pietzenuk B, Seymour DK, Koenig D, Weigel D *et al.* Improving the Annotation of Arabidopsis lyrata Using RNA-Seq Data. *PLoS One* 2015; **10**: e0137391.

143 Gremme G, Brendel V, Sparks ME, Kurtz S. Engineering a software tool for gene structure prediction in higher organisms. *Information and Software Technology* 2005; **47**: 965–978.

144 Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* 2013; **14**: R36.

145 Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* 2011; **29**: 644–652.

146 Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 2012; **28**: 3150–3152.

147 Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 2006; **22**: 1658–1659.

148 Smit AFA, Hubley R. RepeatModeler Open-1.0 http://www.repeatmasker.org. 2008-2015.

149 Smit AFA, Hubley R, Green P. RepeatMasker Open-4.0. . 2013-2015.

150 Bailly-Bechet M, Haudry A, Lerat E. ‘One code to find them all’: a perl tool to conveniently parse RepeatMasker output files. *Mob DNA* 2014; **5**: 13.

151 Lyons E, Pedersen B, Kane J, Freeling M. The value of nonmodel genomes and an example using SynMap within CoGe to dissect the hexaploidy that predates the rosids. *Trop Plant Biol* 2008; **1**: 181–190.

152 Lyons E, Freeling M. How to usefully compare homologous plant genes and chromosomes as DNA sequences. *Plant J* 2008; **53**: 661–673.

153 Rabanal FA, Nizhynska V, Mandáková T, Novikova PY, Lysak MA, Mott R *et al.* Unstable Inheritance of 45S rRNA Genes in Arabidopsis thaliana. *G3* 2017; **7**: 1201–1209.

154 Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013; **29**: 15–21.

155 Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010; **26**: 139–140.

156 Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* 2019; **37**: 907–915.

157 Kuo T, Frith MC, Sese J, Horton P. EAGLE: Explicit Alternative Genome Likelihood Evaluator. BMC Medical Genomics. 2018; **11**. doi:10.1186/s12920-018-0342-1.

158 Alexa A, Rahnenfuhrer J. topGO: enrichment analysis for gene ontology. *R package version* 2010; **2**: 2010.

159 Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat Protoc* 2009; **4**: 1184–1191.

160 Hahne F, LeMeur N, Brinkman RR, Ellis B, Haaland P, Sarkar D *et al.* flowCore: a Bioconductor package for high throughput flow cytometry. *BMC Bioinformatics* 2009; **10**: 106.

161 Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 2011; **27**: 764–770.

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

163 Genomes Consortium. Electronic address, magnus nordborg gmi oeaw ac at, Genomes, Consortium. 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell* 2016; **166**: 481–491.

164 McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010; **20**: 1297–1303.

165 Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* 2012; **6**: 80–92.

166 Mandáková T, Lysak MA. Chromosome Preparation for Cytogenetic Analyses in Arabidopsis. *Curr Protoc Plant Biol* 2016; **1**: 43–51.

167 O'Malley RC, Huang S-SC, Song L, Lewsey MG, Bartlett A, Nery JR *et al.* Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. *Cell* 2016; **165**: 1280–1292.

Chapter 3

Genome size evolution in *Arabidopsis*

Robin Burns[1], **Aleksandra Kornienko[1], Almudena Molla Morales**[1], **Joanna Gunis**[1], **Danièle L. Filiault**[1], **Magnus Nordborg***[1]
[1]Gregor Mendel Institute, Austrian Academy of Sciences, Vienna BioCenter, Vienna, Austria.
***magnus.nordborg@gmi.oeaw.ac.at**

Abstract

Angiosperm genomes vary greatly in size. Evolutionary transitions from outcrossing to selfing have been associated with genome size reduction in *Arabidopsis thaliana* compared to its outcrossing sister species *A. lyrata*. Selection for a compact *A. thaliana* genome likely involved repetitive and intergenic regions such as transposable elements (TEs). Here we use a multi species comparative genomics approach to determine the evolutionary history and the extent of the variation of the genome regions involved in the change in genome size in *A. thaliana*. We aimed to answer the question of whether the length variants are due solely to derived mutations occurring in *A. thaliana* that are fixed in the species or if they could instead represent ancestral polymorphisms that were already present in the ancestor of *A. thaliana* and *A. lyrata,* and that which selection could readily act on. We do not find evidence for selection from standing variation in the ancestor and instead we find that for each variant position in the reference genome of *A. lyrata* multiple alleles are segregating both within and between *Arabidopsis* species, suggesting that these regions of the genome are highly unstable.

Introduction

Angiosperms (flowering plants) vary almost 2000 fold in their genome size. To date, the smallest recorded angiosperm genome belongs to *Genlisea*[1] at 64Mb and the largest to *Paris japonica*[2] at 149Gb, ~ 50 times the size of the human genome and ~3 times that of the largest animal genome sequenced so far[3]. Major processes that lead to genome size expansions are: polyploidization[4] and sometimes repeated cycles of it[5], transposable element (TE) proliferation[4,6–8] and the accumulation of supernumerary ribosome DNA (rDNA) copies[9]. Countering this, several mechanisms leading to genome-shrinkage have been proposed including: the wholesale loss of chromosomes[10], homologous and illegitimate recombination[11,12,13] and different pathways involving double strand break (DSB) repair such as DSB repair by single strand annealing[14].

While the molecular mechanisms underlying these processes are being studied, the nature of the relationship between selection and genome size is less well understood. Much of the variation of genome size has been attributed to varying rates of insertion and deletion[15,16]. Additionally, selection acting on small indels may not always prevent their fixation by genetic drift[17,18], which can also lead to (nearly) neutral variation in genome sizes. However, evidence for selection acting on genome size has been recently reported in maize, where the parallel reduction of genome size with increasing altitude has been linked to flowering time[19]. Other important phenotypes correlated with genome size include: cell size and stomatal density in plants[20], metabolic rates associated with powered flight in birds[16,21] and growth temperatures in bacteria[22]. Correlation with these important phenotypes suggests genome size may play a role in the selection acting on these traits, though it is likely not the causative factor.

The genus *Arabidopsis* holds a wealth of information on the evolution of genome size due to large genomic change over relatively short evolutionary time. The species *A. lyrata* is a predominantly self-incompatible perennial plant with a genome size of roughly 200Mb. 200Mb is the family average of *Arabidopsis*, which are generally obligate outcrossing species. However, the genome of the self-incompatible plant *A. thaliana* has one of the smallest angiosperm genomes at just 125Mb. Since the species divergence ~6 Mya, *A. thaliana* has seemingly reduced its genome size to almost half the ancestral size.

Selfing plants and animals typically have smaller genomes than their outcrossing congeners[23–25]. In plants, the evolution of a smaller genome size is likely due to a reduction in the number of repetitive elements of the genome, such as transposable elements (TEs)[23]. A reduction in the number of TEs in the genome of selfing plants has been hypothesized to be the consequence of exposing recessive deleterious alleles by inbreeding, and their subsequent purging from a population by selection[26]. Evidence for purging of TEs, deleterious SNPs and other repetitive elements has been reported in selfed maize lines where rapid genome shrinkage was seen in just a few generations[27]. Selfing has also been linked to a reduction in TE content of the genome through a lower effective population size (*Ne)* in selfers[28] and the reduced ability of TEs to spread through sex in selfing populations [29]. Genomes of selfers are also likely to experience increased selective pressure to regulate TE mobilization in the genome, as suggested by the lower efficacy of RNA-dependent DNA methylation silencing of TEs in *A. lyrata* compared to *A. thaliana*[30]. However, many theories also support the contrasting hypothesis that selfers will accumulate TEs in the genome. For example, genetic bottlenecks and lower *Ne* could reduce the efficiency of selection to remove deleterious alleles from a population. Homozygous genomes also have a reduced opportunity for ectopic recombination to occur between TE insertions[31]. While differences between outcrossing and selfing *Arabidopsis* have been investigated previously using the reference genomes of *A. thaliana* and *A. lyrata*[23], a comparative genomics approach that examines changes in genome size both within and between different *Arabidopsis* species and mating types is lacking. Such an analysis would shed greater light on how the genomes of selfers and outcrosses vary in their organisation and composition and what selective processes may be involved in genome shrinkage.

To investigate the evolution of genome size in *Arabidopsis*, we therefore collected genome and population data for 4 Arabidopsis species. We also generated 2 additional chromosome-level genomes for *A. lyrata*. We made use of 7 publically available genomes for *A. thaliana*[32] in addition to the *A. thaliana* and *A. lyrata* reference genomes. We used the individual assemblies for *A. halleri*[33] and *A. arenosa* (here the *A. arenosa* subgenome of the allotetraploid *A. suecica*[34]) for a broader sampling of the genus, and because both these genomes have lower heterozygosity than typical *Arabidopsis* outcrossers[35]. This facilitates whole genome alignment and the calling of structural variants.

Through taking the multi species comparative approach, we aimed to answer the plain question — how does a genome shrink? In answering this question we investigated the distribution of positive and negative length variants between the reference genomes of *A. thaliana* and *A. lyrata* in the additional individuals and species of *Arabidopsis*. In contrast to expectations, we find that at the majority of the variant sites multiple length variants (both positive and negative) that are unique to a given genome exist, which suggests that these regions of the genome are unstable. In *A. thaliana* the majority of the multi variant sites are negative length variants, suggesting that these unstable regions may be getting shorter in *A. thaliana* compared to the outcrossing genomes. Sites that are highly polymorphic between *A. thaliana* and the outcrossing genomes constitute a proportion of TE sequences than length variants that are differentially fixed. Purifying selection acting on TE insertions in these genomic regions may therefore be operating similarly between the outcrossers and *A. thaliana*, however in *A. thaliana* selection is likely more efficient. More *A. thaliana* genomes from multiple individuals are required to determine if selection for a compact genome is still ongoing in *A. thaliana*.

Results

Chromosome level assemblies of two additional *A. lyrata* genomes

As only one *A. lyrata* genome (MN47) is publically available for the species, we assembled two additional genomes of *A. lyrata* (“11B02” and “11B21”) from a naturally inbred North American *A. lyrata* population, using 70X Oxford Nanopore Technologies (ONT) MinION long reads (see Methods). We used self compatible accessions of *A. lyrata* as their naturally low levels of heterozygosity greatly assist genome assembly. The assembly size for 11B02 amounted to 217.5Mb with an N50 contig length of 5.0Mb and N50 scaffold length of 27.9Mb, while that of 11B21 amounted to 202.1Mb with an N50 contig length of 8.75Mb and a scaffold N50 of 25.1 Mb (see Supplementary Table 1). The assemblies represent ~85% of the genome size estimated for both the *A. lyrata* accessions using kmer analysis (see Methods). Scaffolds were anchored into 8 pseudo-chromosomes using the reference genomes of *A. lyrata* and the *A. arenosa* subgenome of *A. suecica*[34] in order to correct the described misassembly in the reference (see Supplementary Figure 1).

The assembly statistics are similar to long read PacBio genomes of *A. thaliana*[32] and gene completeness is high with 92% and 95.8% of eudicot BUSCO genes detected in 11B02 and 11B21, respectively. Notably the assemblies constitute fewer stretches of “N” bases and have an overall greater assembly size than the *A. lyrata* reference (see Supplementary Table 1). Gypsy elements make up most of the TE copies in each of the 11B genomes (see Supplementary Figure 2). This agrees with the reference genome of *A. lyrata*[23] and differs from *A. thaliana* where Helitron elements are the most abundant in the genome[36]. The rate of Gypsy transposition in *A. lyrata* has also been previously reported to be 5 times the rate in

A. thaliana[37], highlighting the usefulness of *Arabidopsis* in understanding TE family evolution.

We initially annotated 28,586 and 28,579 protein coding genes for 11B02 and 11B21, respectively (see Methods). This is fewer than the 33,221 protein coding genes previously annotated for the *A. lyrat*a reference[38] and more similar to the 27,445 genes in the A. thaliana and the 26,521 genes in the genome of the outgroup *C. rubella*[39]. However, an analysis of gene orthology (GO) relationships between *Brassicaceae* (see Methods) demonstrated that the 11B02 and 11B21 gene annotation is incomplete, with genes related to immunity and defence as well as genes involved in pollen recognition likely to be missing from the annotation (see Supplementary Table 2). The reduced gene number is likely related to the limited tissue and growth conditions used for the RNA-seq data of the 11B individuals, rather than a misannotation of protein-coding genes in the *A. lyrata* reference. We performed an Interproscan analysis of genes unique (i.e. having no ortholog) to the *A. lyrata* reference (N=1,054) compared to other *Brassicaceae* genomes. If a GO term related to repeat content were overrepresented for the unassigned genes in the *A. lyrata* reference this might reflect that many TEs were missannotated as protein-coding genes. However, GO terms were not overrepresented. We therefore performed lift-over of immunity related genes that are missing in the 11B genomes using the *A. lyrata* reference ortholog (see Methods) resulting in the final number of annotated genes to 30,108 and 30,262 for 11B02 and 11B21, respectively (see Supplementary table 1).

Whole genome alignment between the reference genomes confirms genome shrinkage in *A. thaliana* of repetitive regions

The genome of *A. thaliana* is made up of 5 chromosomes that have undergone large genomics rearrangements that includes 3 chromosomal fusions, and the loss of 3 centromeres, in addition to an evident shrinkage of the ancestral genome size to a current size of approximately 125Mb[23]. The loss of the 3 centromeres likely accounts for only ~10% this size difference while the majority of the difference is explained by thousands of small indels[23]. In contrast, the ancestral genome was likely a karyotype of 8 chromosomes and of a larger genome size (~200 Mb), as this is the genome composition of other diploid *Brassicaceae*, species like *A. lyrata*[23] and *C. rubella*[39] (see Figure 1a). In order to investigate the evolution of the genome size in *A. thaliana* we first aimed to confirm previous results[23] of whole genome alignment between reference genomes of *A. thaliana* and *A. lyrata,* and to characterize the difference in genome size between them.

We called large structural variants (>50bp) from the whole genome alignment using the software Assemblytics[40], and filtered to remove breakpoints containing “N“ bases and for regions that could not be anchored in at least 200bp of uniquely aligned sequences (see Methods). The variants were then categorized into “positive” and “negative” allelic variants in relation to the reference genomes (Figure 1b). We classified ~76Mb of genomic sequence as present in the *A. lyrata* reference genome as positive length variants and present in the *A. thaliana* genome as a negative length variant. The median positive length variant size in the *A. lyrata* reference is 1,039 bp and the mode is 146bp. In contrast, positive length variants in the *A. thaliana* reference (with the *A. lyrata* having the negative length variant) sum to ~20Mb, with a median of 659bp and mode of 158bp. For each bin size, the number of positive length variants in the *A. lyrata* reference exceeds the number of positive length variants in the *A. thaliana* reference, and both are skewed to small sizes - typically lower than 10Kb in size (Figure 1c & d).

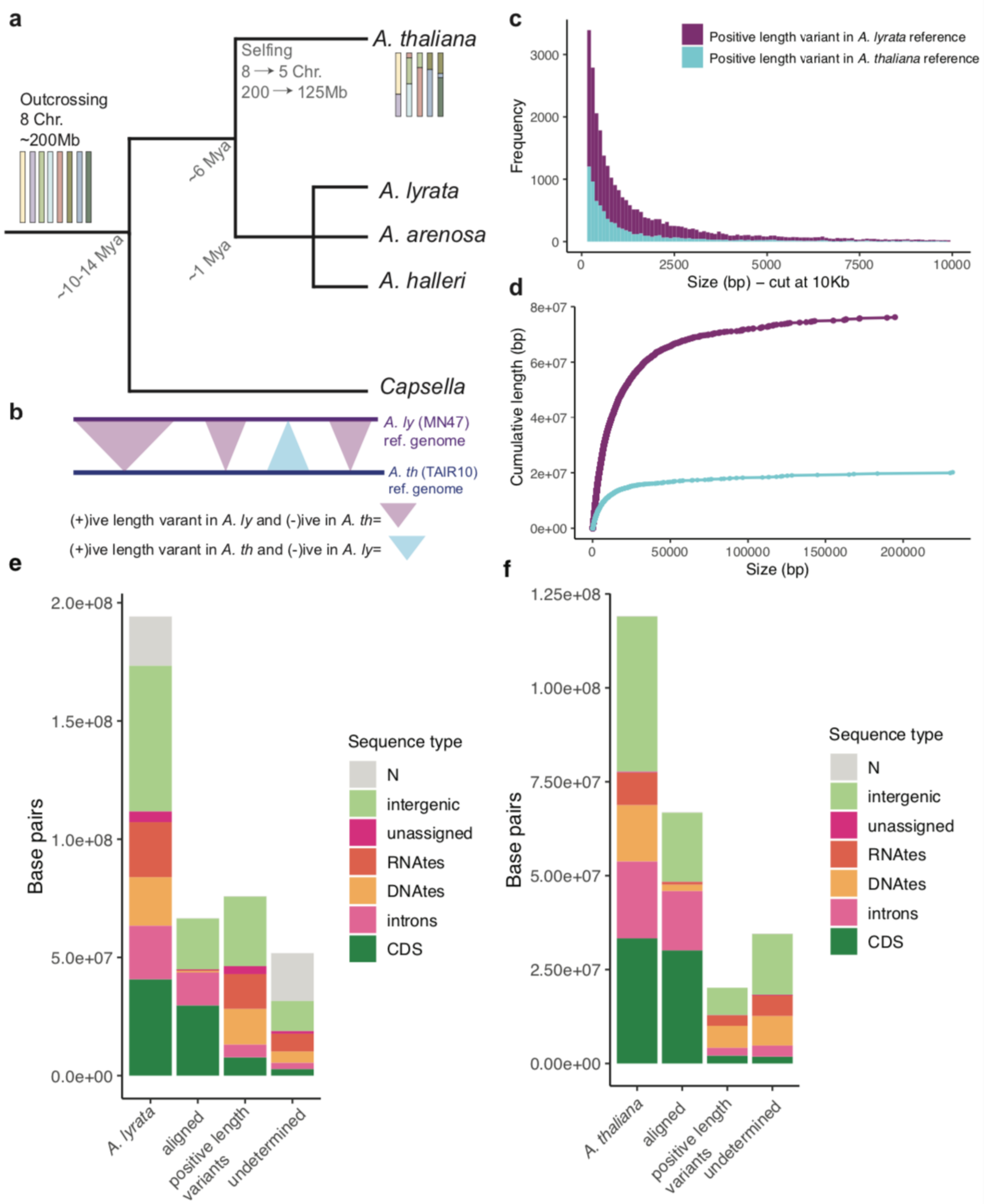


Figure 1 Evolution of genome size in *Arabidopsis*. a The selfer *A. thaliana* has a reduced genome size and karyotype of ~125Mb and 5 chromosomes compared to the ancestral genome size and karyotype of ~200Mb and 8 chromosomes. **b** Whole gene alignment between the reference genomes of *A. thaliana* and *A. lyrata* reveals many length variants between the genomes **c** Histogram of the sizes of positive length variants in each genome. The histogram is cut at 10Kb for display purposes, but maximum sizes of positive length variants can be seen in **d** that is a cumulative plot of the positive length variants in each genome. Sequence composition of the whole genome alignments of **e** the *A. lyrata* and **f** the *A. thaliana* reference genomes.

Roughly ~67 Mb of the two reference genomes align in a one-to-one manner (Figure 1 e and f), and most of these regions are sequences likely to be under evolutionary constraint, such as coding sequences and the introns of protein coding genes. Positive length variants in the *A. lyrata* reference mainly overlap repetitive and intergenic regions, with TE sequences accounting for ~33 Mb of the total ~76Mb of positive length variants. Roughly 51Mb of the *A. lyrata* reference genome could not be determined (i.e. undertermined in Figure 1 e and f) to either a uniquely aligned region or as a positive length variant between the reference genomes. 20 Mb of the undetermined regions contain stretches of N bases, suggesting that some of these regions are poorly assembled. An additional ~21Mb overlaps annotated TE sequences in the *A. lyrata* reference, indicating that some regions are simply too repetitive to be assigned an allele length . Uniquely aligned regions and positive length variants did not show a distinct pattern of enrichment in location along the genome (see Supplementary Figure 3). As the evolutionary history of the undertermined regions is not accessible, we focused our analysis instead on the ~67Mb of uniquely aligned and ~76Mb of positive length variants that occur between the *A. thaliana* and *A. lyrata* reference genomes to investigate genome shrinkage.

Multiple length variants suggest unstable regions in the genomes of *Arabidopsis* are associated with the difference in genome size

Previous work suggested the evolution of the smaller genome size in *A. thaliana* is likely due to numerous derived small deletions occurring along the species branch after the split from its recent common ancestor with *A. lyrata*[23]. This observation was supported by the fact that most of the negative length variants relative to the *A. lyrata* reference are fixed in 95 additional *A. thaliana* accessions (by an analysis of sequenced PCR fragments[41]). This result indicated that selection likely contributed to the smaller genome size of *A. thaliana*[23].

As long reads assemblies for multiple *A. thaliana* accessions are now available, we aimed to re-examine the frequency of the negative *A. thaliana* reference length variant by examining whole genome sequences. One explanation for a difference in genome size between *A. thaliana* and the outcrossing *Arabidopsis* genomes could be that the negative length variant of *A. thaliana* and the positive length variant of *A. lyrata* are differentially fixed between the species groups, such that all *A. thaliana* are fixed for having the *A. thaliana* reference negative length variant and all outcrossers are fixed for having the *A. lyrata* reference positive length variant. Alternatively, the negative length variant in *A. thaliana* may have been segregating in the common ancestor of *A. thaliana* and *A. lyrata.* If a readily available pool of negative and positive length variants existed in the common ancestor, a scenario of selection being able to act on the already existing variation could explain the fast genome size change over relatively short evolutionary time (~6 My). Selection on standing variation would be apparent if the negative length variant of *A. thaliana* is also segregating in the outcrossing genomes at intermediate frequencies. A final alternative explanation for genome shrinkage is that *A. lyrata* and the other outcrossing genomes are expanding in size, though this is unlikely given the outgroup *Capsella* is a similar in genome size and that repetitive sequences are mainly under purifying selection within populations[34,42,43], this hypothesis would be supported however if positive length variants are skewed to high frequencies in the outcrossing *Arabidopsis* genomes.

We examined genomes from additional individuals of *Arabidopsis* from 4 species, *A. thaliana*, *A. lyrata*, *A. halleri* and *A. arenosa*. For *A. thaliana* we downloaded 7 high quality PacBio genomes from[32], and performed a whole genome alignment to the *A. lyrata* reference. For the outcrosser genomes we also aligned the two 11B genomes and the additional genomes of *A. halleri*[33] and *A. arenosa* also to the *A. lyrata* reference (see Methods). We used the *A. arenosa* subgenome of *A. suecica* as a best proxy to *A. arenosa*[34], as the low heterozygosity in selfing *A. suecica* simplifies whole-genome alignment and the calling of structural variants. The genome of *A. halleri* was used as the reference individual was manually inbred for 5 generations and has a heterozygosity rate of 0.04%[33].

We first examined sites where the *A. thaliana* reference has a negative length variant and the *A. lyrata* reference a positive length variant and asked for each additional *Arabidopsis* genome examined which of the 2 variants can be found. From the joint frequency spectrum, we find that the negative length variant of the *A. thaliana* reference is present at high frequencies in the 7 additional *A. thaliana* genomes and is mainly restricted to *A. thaliana*, with very few of the sites segregating in outcrossers. Of the sites that are segregating in the outcrossers, they are mainly low in frequency (Figure 2a). This suggests already that our hypothesis on the negative length allele being selected in *A. thaliana* from existing standing variation in the *Arabidopsis* ancestor does not explain the majority of genome shrinkage. The positive length variant in the *A. lyrata* reference is also mainly restricted to the outcrossing genomes but it is not skewed to high frequencies (Figure 2b), which indicates that the main explanation for genome shrinkage is selection acting on negative length variants in *A. thaliana*, rather than genome expansion in the outcrosser genomes.

The classification of length variants in the additional *Arabidopsis* genomes suggests that the negative length variant in the *A. thaliana* reference genome is a derived mutation that is getting fixed in the species, however, huge variation in the frequency of the *A. thaliana* and *A. lyrata* reference alleles is evident (see Figure 2 a and b and Supplementary Figure 3), and of the ~76Mb difference between the *A. thaliana* and *A. lyrata* reference, only ~740Kb was differentially fixed between the outcrosser and *A. thaliana* genomes. We therefore asked if unique structural variants in the genomes of the *Arabidopsis* individuals overlapped the same genomic regions where the positive and negative length variants between the *A. thaliana* and *A. lyrata* reference exist (see Methods). The existence of multiple length variants segregating at the same genomic location could suggest that the regions of the genome that have shrunk in size between the *A. thaliana* and *A. lyrata* reference are inherently unstable or fragile genomic regions in *Arabidopsis*. We found multiple additional alleles for each site where the reference genomes differ, out of the 22,257 alleles that exist only 2,302 are differentially fixed between the species (see Figure 2c), with *A. thaliana* being fixed for the negative and the outcrossers being fixed for the positive reference length variants. In the 7 additional *A. thaliana* genomes many unique negative length variants exist and overlap the *A. thaliana* reference negative length variant, while in the outcrossing genomes unique positive and negative length variants exist (see Figure 2 c and Supplementary Figure 6). This agrees with the hypothesis that sites where the *A. thaliana* and *A. lyrata* reference genomes differ in size are inherently unstable, and suggests that in *A. thaliana* these sites are mainly getting smaller.

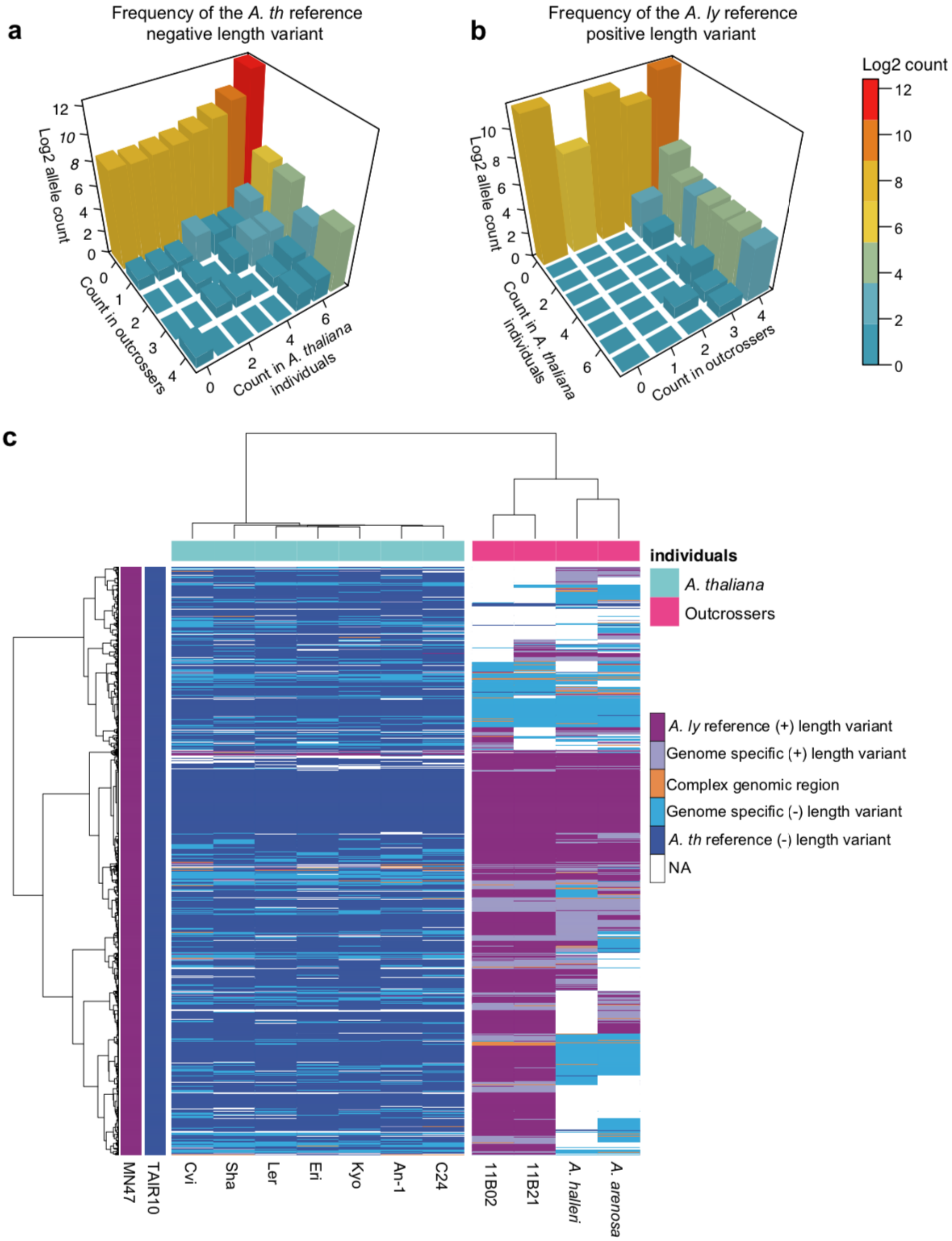


Figure 2 Frequency of the reference genome alleles in additional *Arabidopsis* individuals. a A 3D histogram of joint frequency spectrum for **a** the *A. thaliana* reference or **b** for the *A. lyrata* length variant. **c** A heatmap of length variants found in the different *Arabidopsis* individual and species genomes. If an exact match was found the individual was magenta for the *A. lyrata* reference or dark blue for the *A. thaliana* reference. If no exact match was found but a genome specific length variant overlapped the site the individual was light purple (positive length), light blue (negative length), orange (many variants) or NA (no condition satisfied), using the *A. lyrata* as the reference.

The instability of these regions is also supported by examination of the genome from the outgroup *C. rubella*. If these sites have experienced a single derived change (e.g. either shrinkage in *A. thaliana* or expansion in *A. lyrata*) then we should be able to polarize these sites by examining the genome of *C. rubella*. However, we were unable to polarize the length variants using the genome of *C. rubella* as an outgroup, as just 22 of the negative length variants and 535 of the positive length variants from the *A. thaliana* and *A. lyrata* references, respectively, are present in the *C. rubella* genome. This is expected if these regions of the genome are unstable, and could suggest that the instability goes back further than in *Brassicaceae*. Interestingly, genome specific negative length alleles in *C. rubella* mainly overlapped the variants between the *A. thaliana* and *A. lyrata* reference genomes (see Supplementary figure 6 & 7) and, unlike the outcrossing *Arabidopsis* genomes, very few genome specific positive length variants in *C. rubella* overlapped these sites (1,342 positive compared to 6,133 negative length variants). The genome of *C.* rubella shows a euchromatic profile (i.e. distribution of transposable elements in the genome) that is more similar to *A. thaliana* than to that of *A. lyrata*, despite *C. rubella* having a genome closer in size to *A. lyrata*[39]. This difference is likely explained by the recent transition to selfing in *C. rubella* ~200Kya. However, *C. rubella* may also just be too diverged (MRCA is ~10-14 Mya; see Figure 1a) to accurately call structural variants in these sites, as 14,051 of the 22,257 sites could not be assigned as aligned or overlapping a structural variant (i.e. they were NA calls).

In summary, the sites involved in the genome size difference between the *A. thaliana* and *A. lyrata* reference genomes are not a simple case of a derived mutation and ancestral sequence that are differentially fixed between the species. We also do not find evidence that selection from standing variation contributed to the smaller genome size as the negative length variant in *A. thaliana* is mostly absent from the outcrosser genomes, implying the genome size difference likely involved *de novo* mutations.

Highly polymorphic sites overlap transposon sequences and fixed differences involve introns of highly expressed genes

The multiple alleles segregating in *Arabidopsis* at the length variable sites between the *A. thaliana* and *A. lyrata* reference genomes suggests that these sites are unstable. To characterise these sites, we compared regions that are highly polymorphic and segregating in *Arabidopsis* to regions that are differentially fixed between *A. thaliana* and the outcrosser genomes.

While the number of variants in each group are comparable in number (see Figure 3a), the amount of the *A. lyrata* reference genome sequence explained by each of the groups differ by an order of magnitude. Regions that are differentially fixed between *A. thaliana* and the outcrossers explain the least amount of sequence (~740Kb) while negative length variants that are segregating in both *A. thaliana* and the outcrossers (~8Mb) explain the most (see Figure 3b). In addition, the regions that are differentially fixed are shorter in length than regions that are polymorphic (see Figure 3b).

No distinct genome location difference between the fixed and segregating sites was apparent in the *A. lyrata* reference genome (see Supplementary Figure 8). Examining the makeup of these genomic regions, we find that intergenic sequences explain a high proportion of sequence in each group. Interestingly, length variants that are differentially fixed between *A. thaliana* and the outcrossing genomes are also enriched for introns (~320Kb of the total ~740Kb overlap introns in the *A. lyrata* reference genome), while very few base pairs overlap TE sequences (see Figure 3c). Negative length variants that are polymorphic in the *Arabidopsis* genomes do overlap TEs and the different TE classes (i.e. retrotransposons and DNA elements) together make up the majority of the sequence. Negative length variants segregating in both *A. thaliana* and the outcrosser genomes overlap ~2Mb of retrotransposons and ~1.8Mb overlap DNA TEs (see Figure 3d).

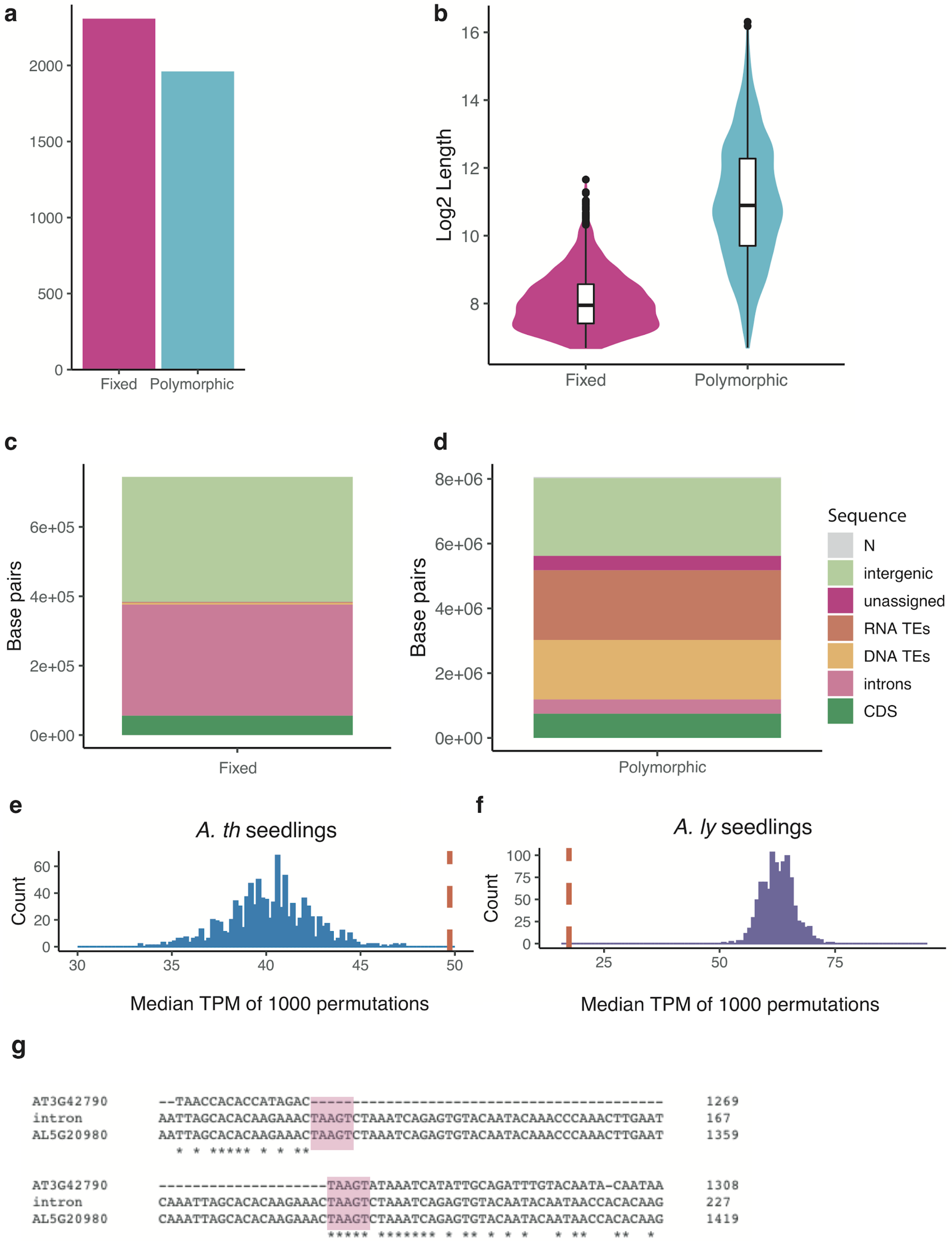
Manually examining 3 local realignments between the *Arabidopsis* genomes for negative length variants segregating in *Arabidopsis,* and that also overlap intact LTR retrotransposons, revealed that the different breakpoints are not located within or around the intact LTR retrotransposons but instead are located further upstream or downstream (often >1kb away). This suggests that the multiallelic negative length variants represent large repeat clusters instead of a simple presence or absence variant of a TE sequence.

Next we turned to fixed differences between the *A. thaliana* and outcrosser individuals, with a focus on length variants that overlap introns, as these variants are likely to have a functional impact on genes. Negative length variants that lead to a reduced intron size or complete loss of an intron in *A. thaliana* compared to the outcrosser genomes matches previous reports[44,45]. 1,136 genes have a positive length variant in the intron of the *A. lyrata* reference genome and a negative length variant in *A. thaliana* reference genome*.* No significant GO enrichment was found for the genes involved. Retrotransposition in *A. thaliana* likely does not explain the intron size difference between the *A. lyrata* and *A. thaliana* orthologs, as not all negative length variants completely overlap an intron and most intron loss occurs at the first intron of the gene (see Supplementary Figure 9 a and b). Genes that have likely retrotransposed in *A. thaliana* (i.e. genes that do not have any introns according to the annotation in Araport11) also do not overlap the 1,136 genes.

Interestingly, we find that the length variant between the orthologs is frequently located in the first intron of the gene (see Supplementary Figure 9). Regulatory elements in introns are frequently located in the first intron of a gene, and furthermore genes that are regulated by introns are often highly and broadly expressed[46]. We therefore examined gene expression in *A. thaliana* and *A. lyrata* (see Methods). 691 of the 1, 136 showed evidence of expression in both *A. thaliana* and *A. lyrata* (see Methods). In *A. thaliana* the median expression level of the 691 genes exceeds the median expression for 1000 permutations across four different tissues - seedlings, pollen, roots and flowers. This suggests that these genes are among the highest expressed genes in *A. thaliana*. High gene expression was not observed in *A. lyrata* in seedlings, where instead these genes appeared in the lower tail of expression compared to 1000 permutations of background genes (see Methods). While it is difficult to compare gene expression between diverged species, investigations using F1 hybrids between *A. thaliana* and *A. lyrata* could help answer if the change in intron sequence affected the expression of these genes in cis, while controlling for differences in their trans regulation[47].

Manually examining local realignments for 60 of the 691 genes we find 4 examples where indels in the intron sequence of *A. lyrata* were likely caused by non homologous end joining (NHEJ) of a DSB (see Figure 3X and Supplementary Figure 10 for more examples). The error prone repair of DSBs by NHEJ has been linked to the loss and gain of introns[48]. Another source of intron gains is TE insertions[49] however we did not find many examples of introns with similarity to TE sequences (just one example — the first intron of AL8G18870). Although we cannot conclude if DSB repair led to a deletion in the intron in *A. thaliana* or insertion in the outcrosser genomes, we note that a bias in DSB repair could be a factor in genome size evolution. A comparison of DSB repair between barley and *A. thaliana* found the DSBs are preferentially repaired by NHEJ over homologous recombination (HR), which was not the case in barley[14]. Given the high levels of expression for the genes involved, DSBs resulting from transcription may have been repaired by NHEJ in *A. thaliana*. A comparative analysis of DSBs between and within *Arabidopsis* species could shed light on whether *A. thaliana* more often repairs DSBs by NHEJ than HR compared to its outcrossing relatives.

In summary, polymorphic length variants in *Arabidopsis* seem to overlap repeat clusters and fixed length variants overlap introns of highly expressed genes in *A. thaliana*. Finally, NHEJ repair of DSBs may be a factor in generating indels involved in the genome size difference between *A. thaliana* and the outcrosser genomes.



```
AT3G42790       --TAACCACACCATAGAC------------------------------------------    1269
intron          AATTAGCACACAAGAAACTAAGTCTAAATCAGAGTGTACAATACAAACCCAAACTTGAAT    167
AL5G20980       AATTAGCACACAAGAAACTAAGTCTAAATCAGAGTGTACAATACAAACCCAAACTTGAAT    1359
                  * * ***** * * **

AT3G42790       --------------------TAAGTATAAATCATATTGCAGATTTGTACAATA-CAATAA    1308
intron          CAAATTAGCACACAAGAAACTAAGTCTAAATCAGAGTGTACAATACAATAACCACACAAG    227
AL5G20980       CAAATTAGCACACAAGAAACTAAGTCTAAATCAGAGTGTACAATACAATAACCACACAAG    1419
                                    ***** ******* * ** * * *   * **   **  *
```

Figure 3 Characterising differentially fixed and polymorphic length variants in *Arabidopsis*. a A bar plot **b** and a violin plot of the different length variants **c** and **d** show the sequence composition of each variant type. Gene expression for genes that have a fixed negative length variant in *A. thaliana* that corresponds to the first intron of the ortholog in *A. lyrata* for **e** *A. thaliana* and **f** *A. lyrata* seedlings. The orange bar represents median expression of the 691 genes and the blue and purple histograms represent 1,000 permutations using the background genes. **g** An example of an indel occurring in the first intron of *A. lyrata* with 5 nucleotides that show microhomology at the breakpoints that suggest NHEJ repair of DSBs.

Conclusion

We examined the process of genome shrinkage by taking a comparative genomics approach utilizing multiple species and individuals of *Arabidopsis*. We found no evidence that selection from standing variation played a substantial role in the evolution of the smaller genome size in *A. thaliana*. We do find evidence for an enormous amount of variation at the length variable sites between the *A. thaliana* and *A. lyrata* reference genomes that suggests these regions may be unstable. The large amount of the *A. lyrata* reference genome that is segregating as negative length variants in other *Arabidopsis* genomes also brings into question how much of the non-coding genome is dispensable. The overlap of the highly polymorphic sites with large repeat clusters also raises mechanistic questions of how these regions of the genomes accumulate and/or dispense of repeats. The enrichment of TEs in regions of the genome that appear unstable also raises the question of whether TE insertions lead to these regions becoming unstable or whether TEs preferentially insert into already unstable sites of the genome.

Our finding of the shortening of introns in *A. thaliana* being linked to differential repair of DSBs, suggests that biases in the types of DSB repair used play a role in the evolution of genome size in agreement with previous studies on *A. thaliana*[14] and *Genlisea*[50]. The connection with highly expressed genes in *A. thaliana* may be related to transcription cost as suggested by other studies[51] but requires further investigation to determine if it is a cause of consequence. A comparative approach into biases between repair pathways of DSBs used in *Arabidopsis* could shed light on this, in addition to generating CRISPR mutants for indels in these introns and their effect on gene expression in *A. thaliana* and *A. lyrata*.

The high amount of polymorphism makes it difficult to assess the ancestral state of the length variants. This may suggest the outgroups are too diverged to correctly assign an ancestral allele, and indeed this is complicated by the fact that these genomic regions appear unstable even within a species. Future studies like the 1001G+ (https://1001genomes.org/) that aims to generate high quality long read assemblies for many *A. thaliana* accessions will likely allow for a better understanding of the dynamics involved in the selection for a smaller genome size, due to the lower levels of variation within *A. thaliana*. A similar approach could be done for closely related populations of *A. lyrata* in North America, due to the lower genetic divergence for these *A. lyrata* individuals. Examining large numbers of closely related samples could allow for a better understanding of the types of mutations involved and how selection acts on non-coding regions of the genomes, this would provide a good a priori knowledge when comparing genomes between species to understand evolutionary processes over longer evolutionary time.

Materials & Methods

Minion sequencing of 11B *A. lyrata*

We sequenced two North American *A. lyrata* accessions, 11B02 and 11B21. Both individuals come from the 11B population of *A. lyrata*, which is self compatible and situated in Missouri [52] (GPS coordinates 38° 28' 07.1" N; 90° 42' 34.3" W) . Plants were bulked for 1 generation in the lab and DNA was extracted from ~20g of 3 week old seedlings, grown at 21°C and dark treated for 3 days prior to tissue collection. DNA was extracted using a modified protocol for high molecular weight DNA extraction from plant tissue. DNA quality was assessed with a Qubit fluorometer and a Nanodrop analysis. The final output of MinION sequencing for 11B02 was 13,67 Gbp in 763,800 reads and an N50 of 31,15 Kb. The final output of MinION sequencing for 11B21 was 17.55 Gb, 1.11 M reads with an N50 of 33.26 Kb.

Genome assembly, polishing and scaffolding

We assembled the genome of the two *A. lyrata* accessions 11B02 and 11B21 using Canu[53] (v 1.8) with default settings and a genome size set to 200Mb. The genome of 11B02 is contained in 498 contigs and the genome of 11B02 in 265 contigs. The contig assemblies were polished using Racon[54] (v 1.4) and ONT long reads were mapped using nglmr[55] (v 0.2.7). Assemblies were further polished by mapping PCR-free Illumina 150bp short reads (~100X for 11B02 and ~88X for 11B21) to the long read corrected assemblies. Short read correction of assembly errors was carried out using Pilon[56] (v1.23). Contigs were scaffolded into pseudo-chromosomes using Ragoo[57] and by using the error corrected long reads from Canu and the *A. lyrata* reference genome and the *A. arenosa* subgenome of *A. suecica* as a guide followed by manual inspection of regions. The assembly size for 11B02 was 213Mb and 11B21 was 202Mb. Genome size was estimated using findGSE[58] with a resulting estimated genome size of ~256Mb for 11B02 and ~237Mb for 11B21.

RNA extraction and sequencing

RNA was extracted from 3 week old seedlings grown at 21°C from the same batch of plants used for the DNA extraction for long read sequencing (above), without dark treatment. Total RNA was extracted from whole rosettes using the ZR Plant RNA MiniPrepTM kit and libraries were prepared using the NEBNext Ultra II RNA Library Prep Kit for Illumina. The libraries were PCR amplified for 7 cycles. 150bp paired-end sequencing was carried out at the VBCF on Illumina (HiSeq 2500) using multiplexing, The total number of paired-end reads amounted to 23.364 M reads for 11B02 and 24.478 M reads for 11B21.

Genome annotation

We used RepeatModeler2[59] to build a *de novo* consensus library of TE sequences and classify TE sequences into TE families, with the added LTR pipeline extensions of RepeatModeler2. TEs of “Unknown” sequences were blasted to the TE sequences annotated in the *A. thaliana* reference genome and TE sequences annotated for the reference *A.lyrata* reference genome. TEs with >80% identity for >80% alignment were assigned to a TE family, while TEs without a good BLAST hit were labelled as TEs of "Unknown" sequences. Genome locations of the identified TE consensus sequences were determined by using RepeatMasker[60] (version 4.0.7) and filtered for full length matches using a code described in Bailly-Bechet et. al[61].

We combined *ab initio*, homology and transcriptome based approaches to predict protein-coding genes. For *ab intio* based prediction we trained AUGUSTUS[62] (v3.2.3) on a set of single copy conserved genes from eudicotyledons (eudicotyledons_odb10) using BUSCO[63] (v 4.0.6) . For homology based prediction we used protein sequences from the reference *A. lyrata* genome from the second version of *A. lyrata* annotation[38] (Alyrata_384_v2.1) and aligned the protein sequences using GenomeThreader[64] (v 1.7.0) to the long read assemblies of *A. lyrata* in order to annotate gene structures. For a transcriptome based approach, we aligned the 150bp PE reads of RNAseq (above) to a repeat masked version of the *A. lyrata* genome assemblies using STAR[65] (v.2.7.5) to identify exon regions and splice junctions. The alignment results were then assembled into transcriptomes using Trinity[66] (v 2.9.1). Assembled transcripts were filtered for a TPM of 0.5, and the longest reading frame was chosen using Transdecoder[67] (v 5.5.0). Protein sequences from the assembled transcripts were then aligned back to the respective genome assemblies using GenomeThreader (v 1.7.0).

To determine the orthology relationship between genes in the different genomes, the software Orthofinder[68] was used. The genomes of the reference *A. thaliana* (TAIR10) and *A. lyrata* (MN47), 7 *A. thaliana* genomes[32] , *A. halleri*[33] and the genomes of *C. rubella* and *C. grandiflora*[39]. Genes that could not be assigned to an orthogroup were denoted as unassigned. Genes that were not present in 11B but were present in the *A. lyrata* reference and *A. thaliana* reference were extracted, and gene ontology was conducted using TopGO[69] using the *A. thaliana* reference ortholog of the missing genes. Genes that had an orthogroup between the *A. lyrata* reference and *A. thaliana* reference were used as the background. Lifover of genes missing in the 11B genomes (but present in *A. lyrata* reference and *A. thaliana* reference) was carried out using the software Liftoff[70].

Identifying structural variants from whole genome alignments

We aligned the reference genome of *A. thaliana* to the reference genome of *A. lyrata* using the nucmer command from MUMmer4[71] with the default settings. Large structural variants were called using the software Assemblytics[40] with the parameters '200 50 1000000', for a unique sequence anchor length of 200bp, a minimum SV size of 50bp and a maximum SV size of 1Mb. Assemblytics outputs SVs as falling into 6 groups: "repeat contraction", "repeat expansion", "deletion", "insertion", "tandem contraction" and "tandem expansion"[40]. From the output of Assemblytics we grouped "repeat-" or tandem contraction" and "deletion" together to represent positive length variants present in the *A. lyrata* reference and the corresponding negative length variant being present in the *A. thaliana* reference. We did not use the "repeat-" or "tandem expansion" and "insertion" output of the SV calls to describe positive length variants in the *A. thaliana* reference and negative length variants in the *A. lyrata* reference. Instead, we performed the reciprocal alignment with the *A. thaliana* reference as the reference genome provided to nucmer, and the *A. lyrata* reference as the query genome. From the reciprocal alignment we grouped "repeat-" or "tandem contraction" and "deletion" as regions where the *A. thaliana* reference has the positive length variant and the *A. lyrata* reference has the negative length variant. We performed the reciprocal alignment to avoid reference bias in SV calls. Reference biases in the ability to find variants where the allele is shorter than the reference genome compared to alleles that are longer than the reference genome have been noted previously[9], and are likely due to alignment algorithms being better at dealing with gaps than insertion sequences[9].

Whole genome alignments were carried out in the same manner to describe long and negative length variant variants between the genomes of the additional individuals and the genome of *A. lyrata* reference. As the *A. lyrata* reference genome has a high proportion of ”N” bases which make the true alignment of a region difficult to interpret, structural variants that contained N bases in the first 50 bases or the last 50 bases of the allele, or made up more than 25% of total bases for a given structural variant were discarded from the analysis. Regions where the *A. lyrata* reference has the positive length variant and query genomes have the negative length variant were combined into a matrix. Negative length variants that had the same breakpoint sites were defined as shared, otherwise the negative length variant was unique to a query genome.

Classifying the *A. lyrata* reference and *A. thaliana* reference length variants in the additional *Arabidopsis* genomes

We classified each individual as having the *A. thaliana* reference negative length variant or the *A. lyrata* reference positive length variant. An individual had the *A. thaliana* reference negative length variant if breakpoints between the negative length variant calls matched the breakpoints of the *A. thaliana* reference negative length variant. If the breakpoints of the *A. thaliana* reference negative length variant were instead contained within a 1-to-1 alignment with the *A. lyrata* reference genome, the individual had the *A. lyrata* reference positive length variant.

If neither condition was satisfied, the breakpoints unique to an individual were queried for overlap with the *A. thalian*a reference negative length variant. If a negative length variant in the genome of a given individual overlapped by 50% reciprocally with the *A. thaliana* reference negative length variant, this was described as an overlapping negative length variant. If a positive length variant of a given individual's genome was contained within the *A. thaliana* reference negative length variant by 50%, this was described as overlap with a positive length variant. If multiple negative and/or positive length variants overlapped the *A. thaliana* negative length variant the region was described as complex. If these conditions were not satisfied, the region was described as “NA”. NA alleles are likely the result of genomic regions that are too repetitive in the whole genome alignment, and do not have a unique sequence to anchor the alignment to. Due to their repetitiveness, the history of the region cannot be concluded. Another reason why a region may be “NA” is an incomplete query genome, where alignments to the *A. lyrata* reference are simply not possible.

Expression analysis

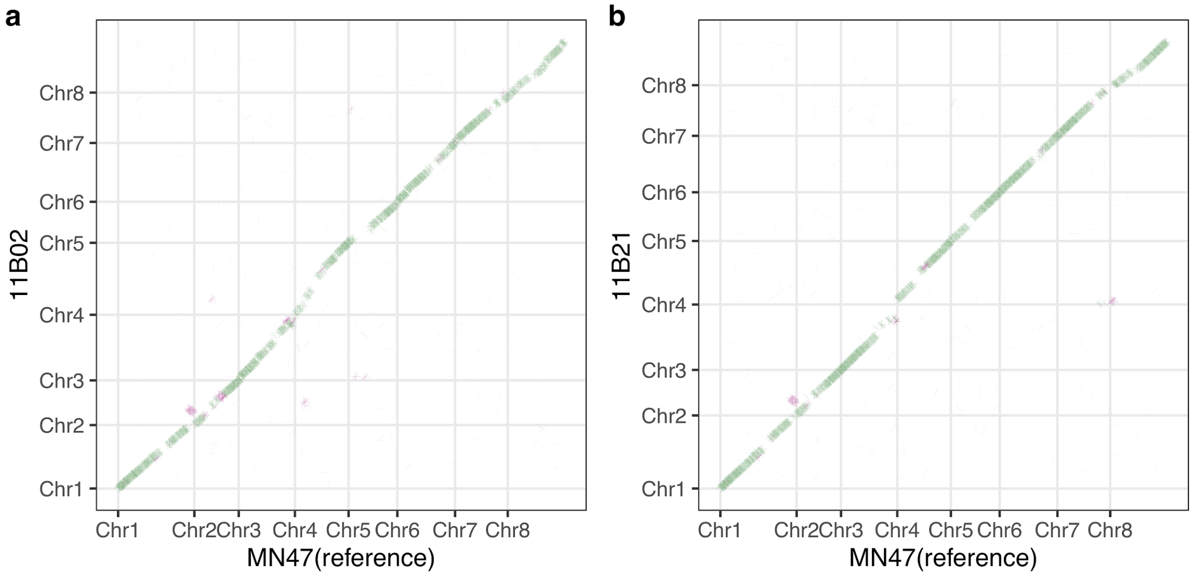
We used expression data from Kornienko 2021 (in progress), with RNA sequencing of *A. thaliana* accessions grown at 21°C with 3 replicates for seedlings, flowers, roots and pollen.

RNAseq data was available for 2 of the the 8 *A. thaliana* accessions examined - Sha and the *A. thaliana* reference TAIR10. Seedlings and roots were of the 9 true leaf stage. Transcripts per million (TPM) were calculated for each gene. RNAseq data for *A. lyrata* was used using the data generated for genome annotation. As a background 1:1 orthologs between *A. thaliana* and *A. lyrata* were used that showed evidence of gene expression (TPM > 0.5) in at least one individual in both *A. thaliana* and *A. lyrata* and across the different tissues examined for *A. thaliana* (seedlings, roots, floral buds and pollen).

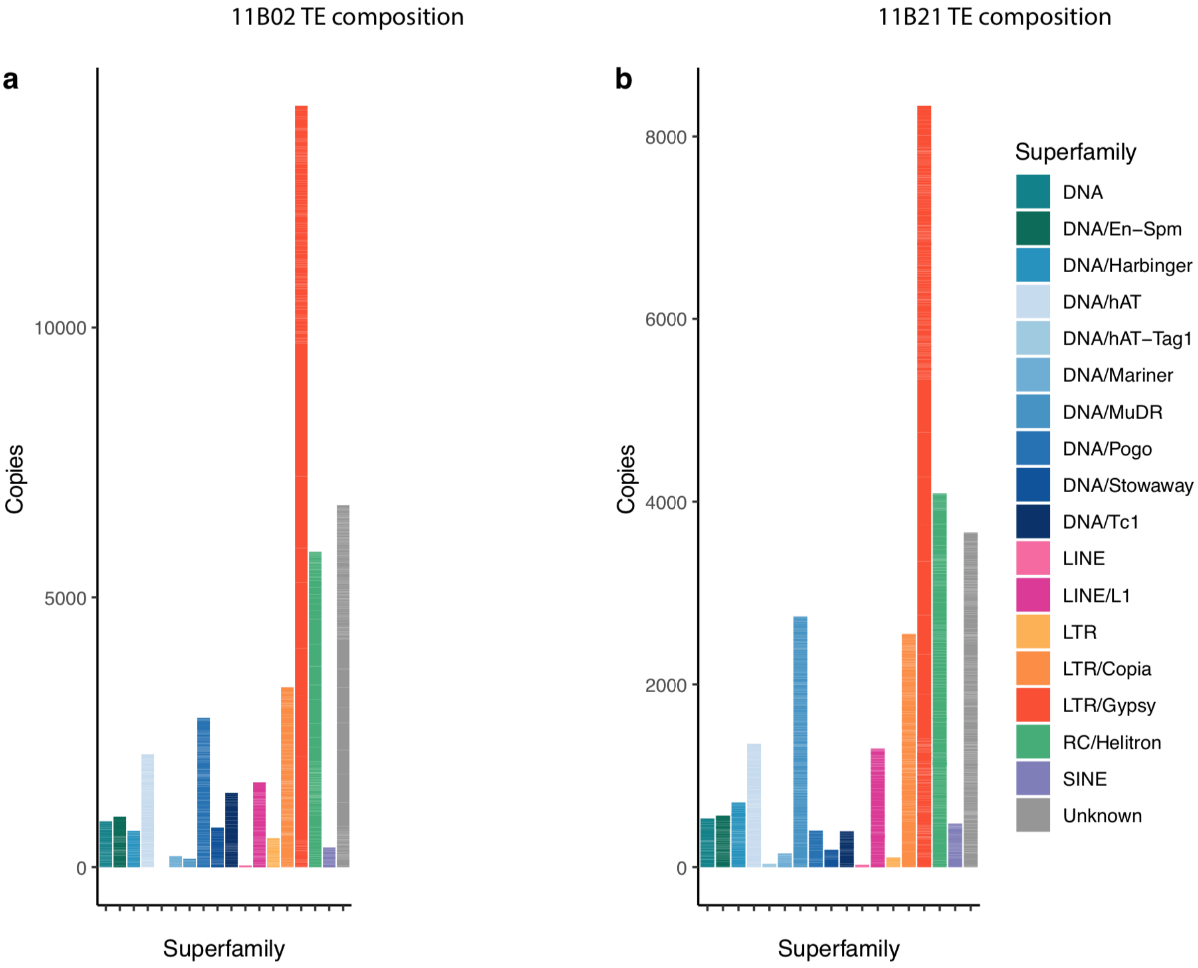
Acknowledgements

We thank Yvonne Willi for providing seeds for the 11B individuals and Shohei Takuno for discussions.

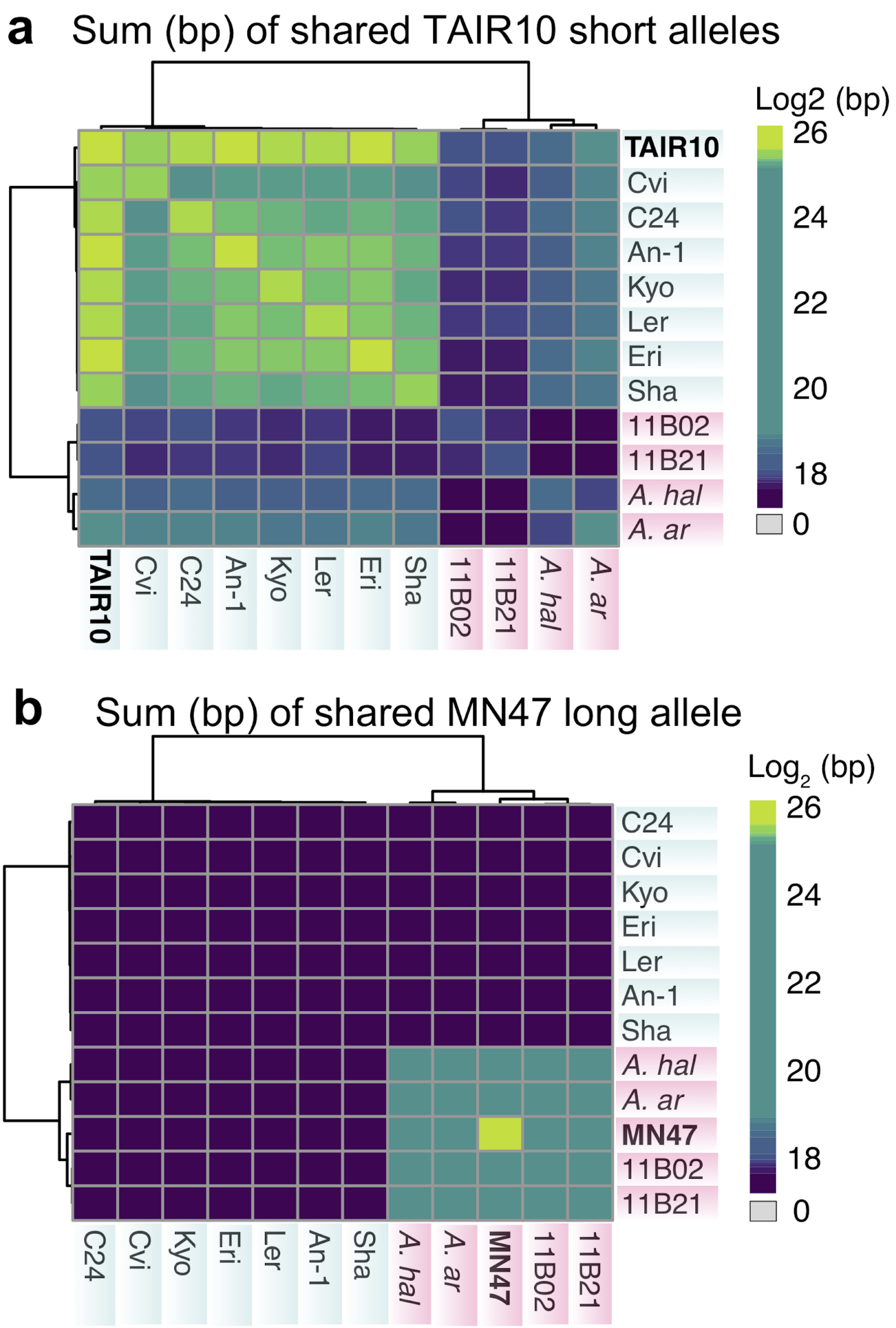
Supplementary material



Supplementary Figure 1. Dot plots of the newly assembled *A. lyrata* genomes (y-axis) to the reference *A. lyrata* genome (A. lyrata reference; x-axis). **a** For 11B02 and **b** for 11B21 the described misassembly in the reference in *A. lyrata* is visible as a translocation (in pink) from Chr1 in A. lyrata reference to Chr2 in both 11B02 and 11B21.



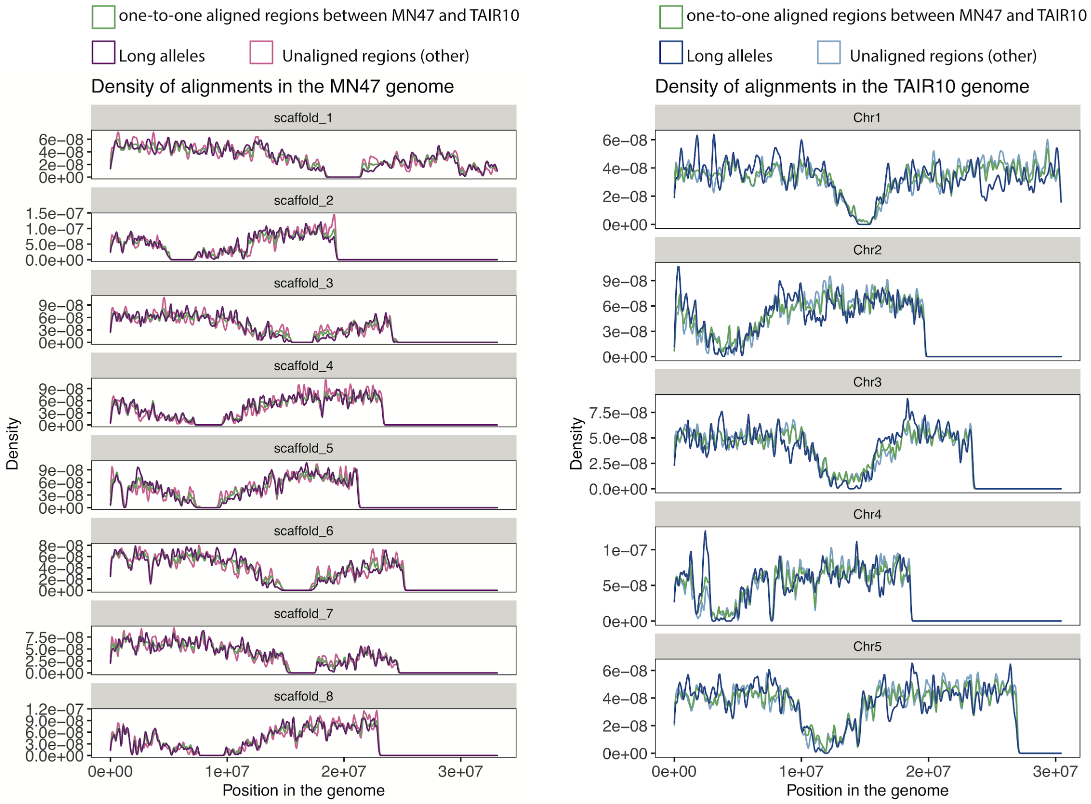
Supplementary Figure 2. TE composition of the 11B genomes show LTR Gypsy elements are the most abundant. a The 11B02 and **b** the 11B21 genome.



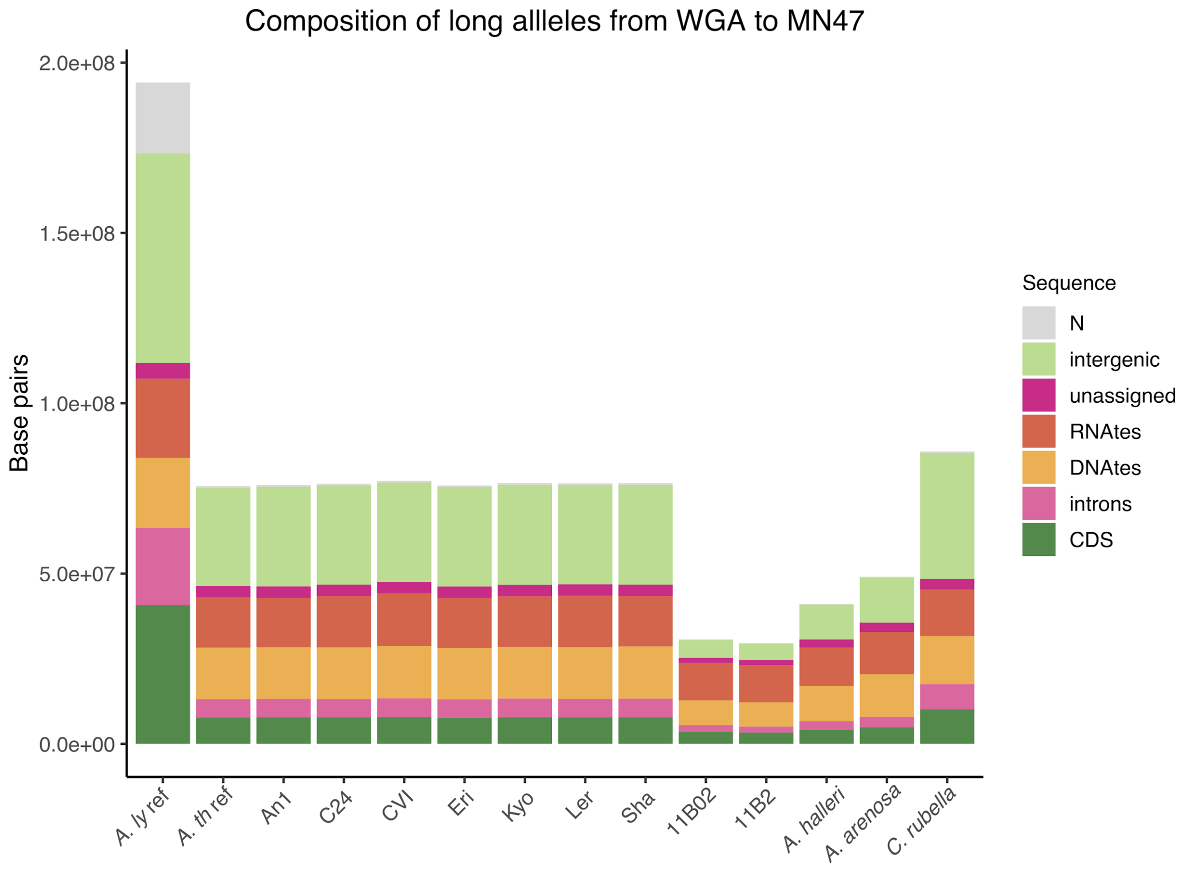
a Sum (bp) of shared TAIR10 short alleles

b Sum (bp) of shared MN47 long allele

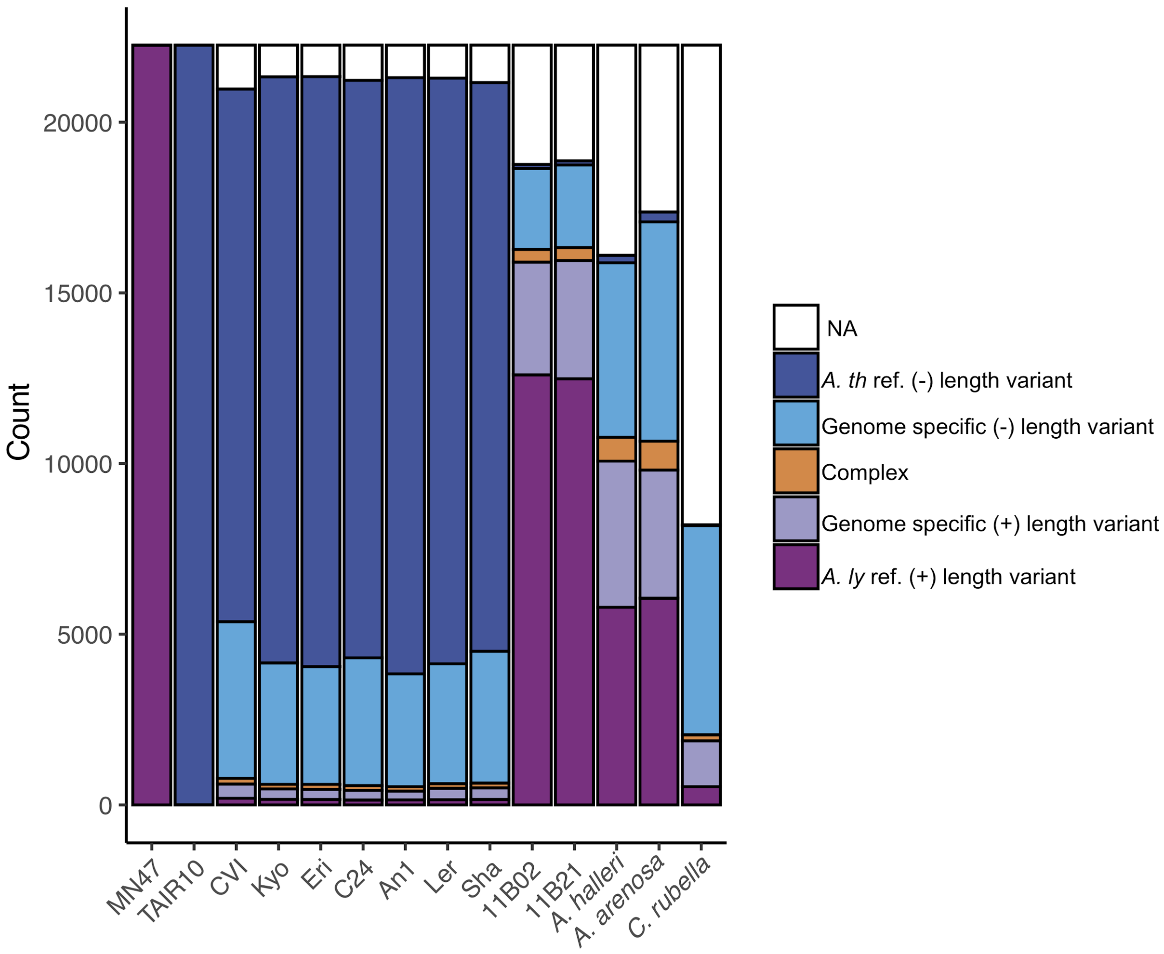
Supplementary Figure 3. Base pairs explained by sharing of the reference variants a reference *A. thaliana* negative length variant or **b** *A. lyrata* positive length variant



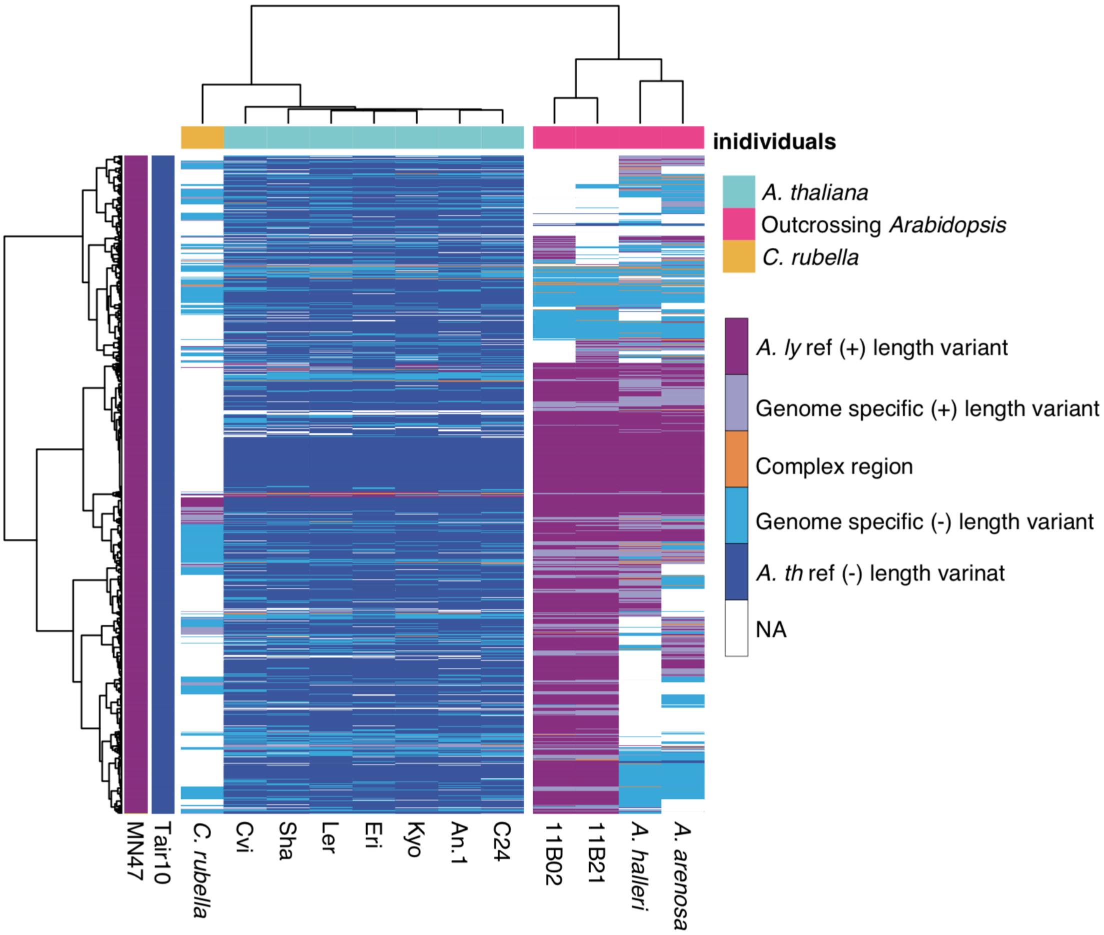
Supplementary Figure 4. Genome-wide distributions of the whole-genome alignment classes between A. lyrata reference and A. thaliana reference from Supplementary Figure 2 for a. The *A. lyrata* reference genome and **b** the *A. thaliana* reference genome.



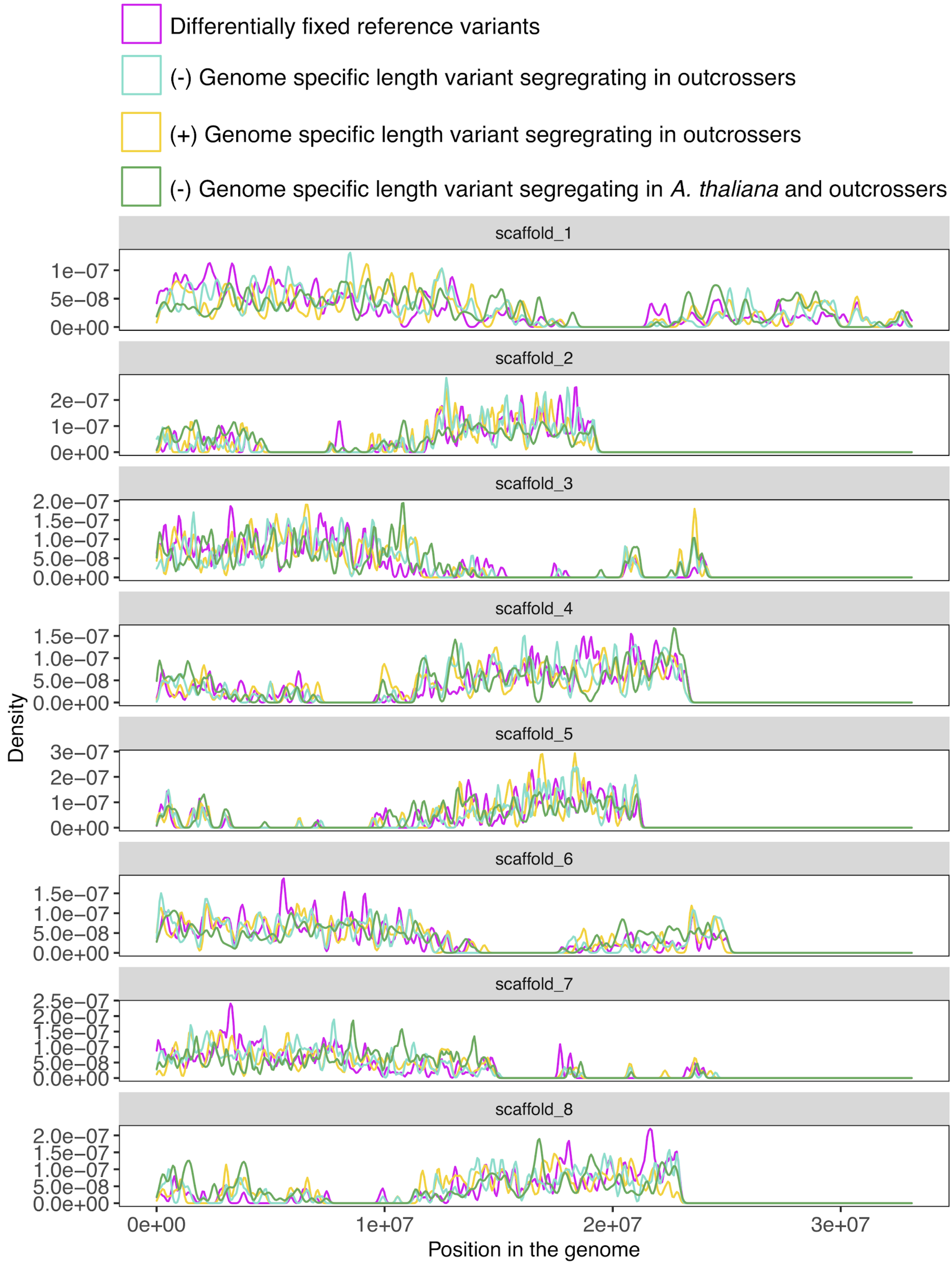
Supplementary Figure 5. Whole genome alignment to the *A. lyrata* reference. Composition of positive length variants from the genome alignments between the *A. lyrata* reference and all *Arabidopsis* individuals and the *C. rubella* genome. The *A. arenosa* genome refers to the *A. arenosa* subgenome of *A. suecica*.



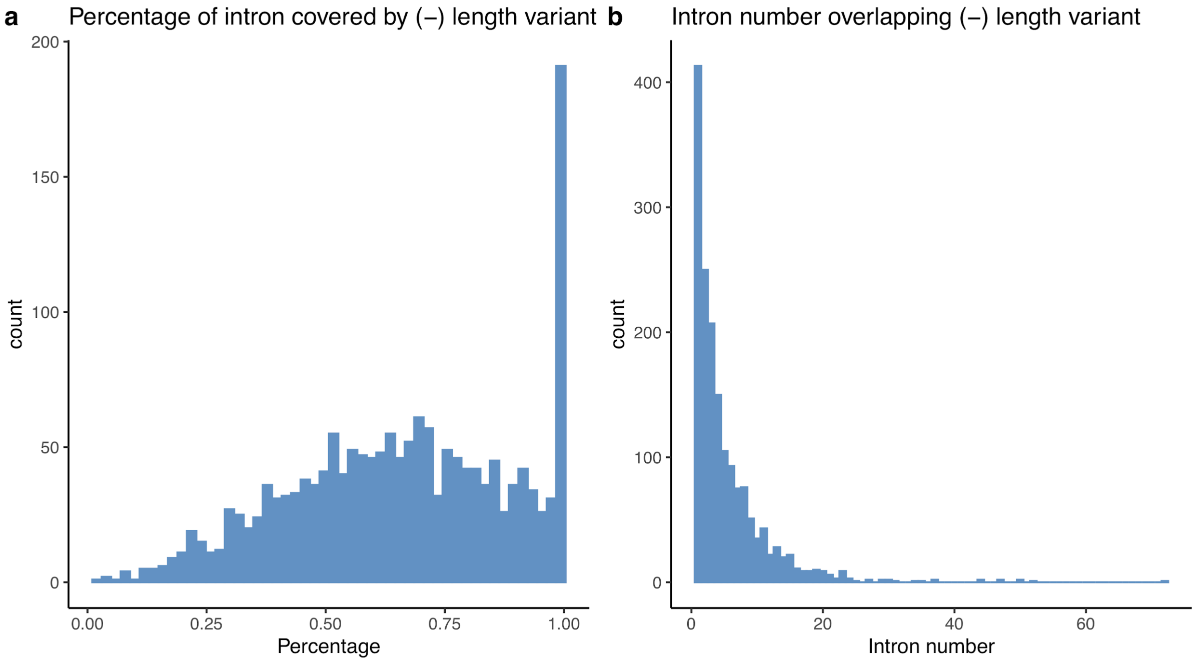
Supplementary Figure 6. Count of structural variants in the analyzed genomes. A breakdown of the 22,257 biallelic sites between the *A. thaliana* reference and *A. lyrata* reference (where the *A. thaliana* reference has the negative length variant and the *A. lyrata* reference the positive length variant) in the additional genomes analyzed.



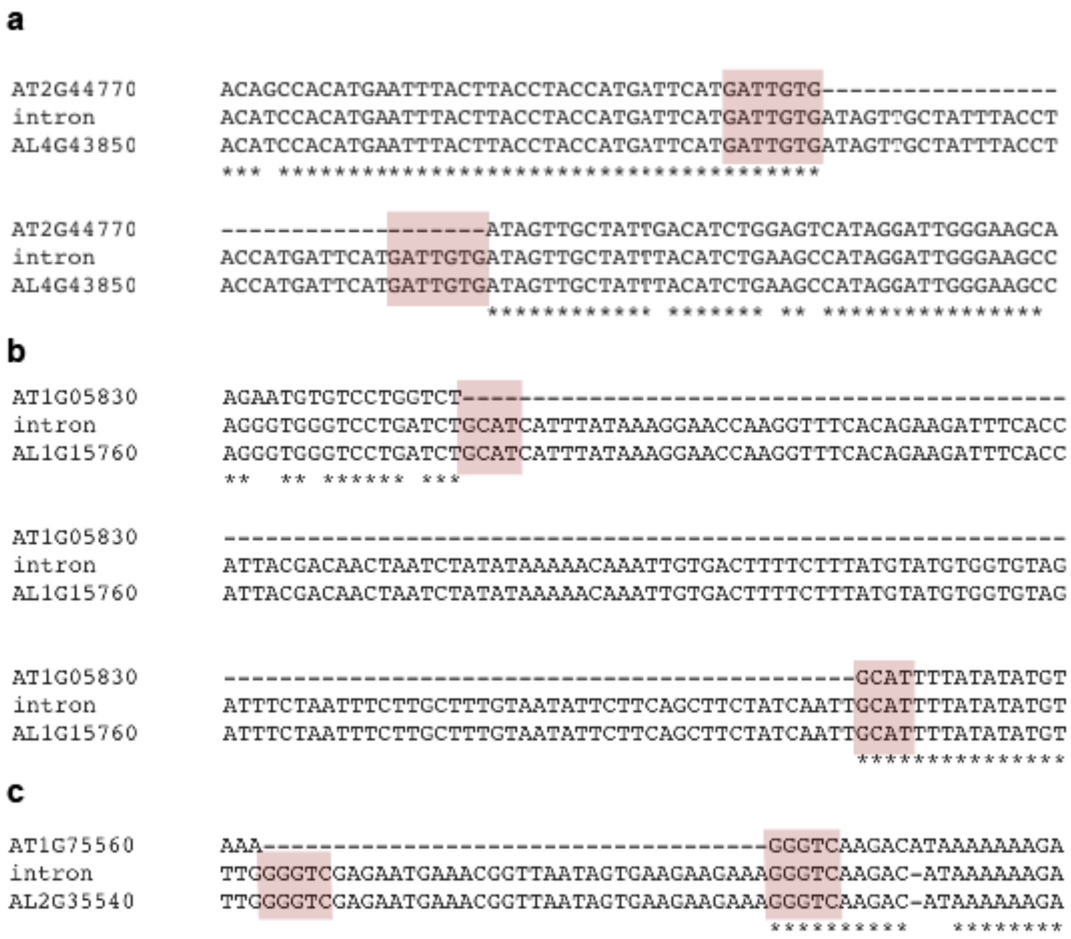
Supplementary Figure 7. Heatmap of biallelic sites between A. lyrata reference and A. thaliana reference. *A. thaliana* reference has the negative length variant and *A. lyrata* reference has the positive length variant. Sites from *C. rubella* mainly can not be called (NA) as the genome may be too diverged or regions of the *C. rubella* assembly may be incomplete (the genome of C. rubella is ~60% complete[39]).



Supplementary Figure 8. Genomic distribution of called length variants.

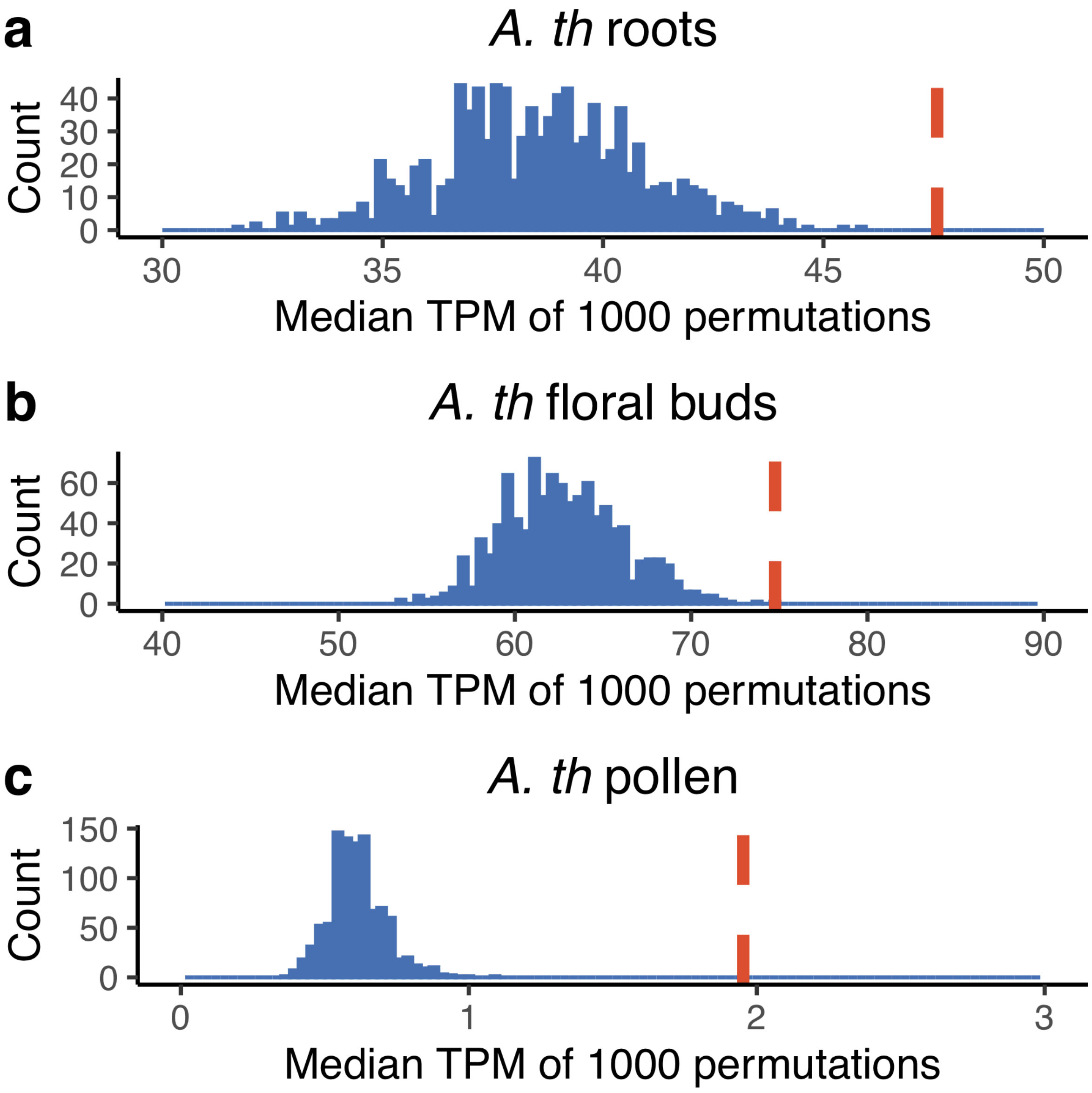


Supplementary Figure 9. Differentially fixed length variants between *A. thaliana* and the outcrossers linked with a possible loss of introns in *A. thaliana* orthologs. a Amount of an intron in the *A. lyrata* reference genome overlapping a negative length variant in the *A. thaliana* reference genome **b** The position of the intron in the gene where the negative length variant occurs.



```
a

AT2G44770       ACAGCCACATGAATTTACTTACCTACCATGATTCATGATTGTG-----------------
intron          ACATCCACATGAATTTACTTACCTACCATGATTCATGATTGTGATAGTTGCTATTTACCT
AL4G43850       ACATCCACATGAATTTACTTACCTACCATGATTCATGATTGTGATAGTTGCTATTTACCT
                *** ***************************************

AT2G44770       -------------------ATAGTTGCTATTGACATCTGGAGTCATAGGATTGGGAAGCA
intron          ACCATGATTCATGATTGTGATAGTTGCTATTTACATCTGAAGCCATAGGATTGGGAAGCC
AL4G43850       ACCATGATTCATGATTGTGATAGTTGCTATTTACATCTGAAGCCATAGGATTGGGAAGCC
                                   ************ ******* ** ****************

b
AT1G05830       AGAATGTGTCCTGGTCT-------------------------------------------
intron          AGGGTGGGTCCTGATCTGCATCATTTATAAAGGAACCAAGGTTTCACAGAAGATTTCACC
AL1G15760       AGGGTGGGTCCTGATCTGCATCATTTATAAAGGAACCAAGGTTTCACAGAAGATTTCACC
                **  ** ****** ***

AT1G05830       ------------------------------------------------------------
intron          ATTACGACAACTAATCTATATAAAAACAAATTGTGACTTTTCTTTATGTATGTGGTGTAG
AL1G15760       ATTACGACAACTAATCTATATAAAAACAAATTGTGACTTTTCTTTATGTATGTGGTGTAG


AT1G05830       ---------------------------------------------GCATTTTATATATGT
intron          ATTTCTAATTTCTTGCTTTGTAATATTCTTCAGCTTCTATCAATTGCATTTTATATATGT
AL1G15760       ATTTCTAATTTCTTGCTTTGTAATATTCTTCAGCTTCTATCAATTGCATTTTATATATGT
                                                             ***************
c
AT1G75560       AAA------------------------------------GGGTCAAGACATAAAAAAAGA
intron          TTGGGGTCGAGAATGAAACGGTTAATAGTGAAGAAGAAAGGGTCAAGAC-ATAAAAAAGA
AL2G35540       TTGGGGTCGAGAATGAAACGGTTAATAGTGAAGAAGAAAGGGTCAAGAC-ATAAAAAAGA
                                                       **********   ********
```

Supplementary Figure 10. Examples of intron indels for 3 examples a, b and c. Mirohomology is indicated in red with the *A. thaliana* ortholog on top, *A. lyrata* ortholog on the bottom and intron of *A. lyrata* in between the genes. Asterisks indicate conserved nucleotides.



Supplementary Figure 11 Analysis of gene expression in different tissues of *A. thalaian*. **a** In roots **b** in floral buds and **c** in pollen for the 619 genes that have a differentially fixed negative length variant in *A. thaliana* that corresponds tp the first intron of the *A. lyrata* ortholog. The orange dashed line represents the median expression of the 619 (in TPM) and the blue histograms represent 1000 permutations of 10, 287 background genes (1:1 orthologs that show evidence of expression in *A. thaliana* and *A. lyrata*)

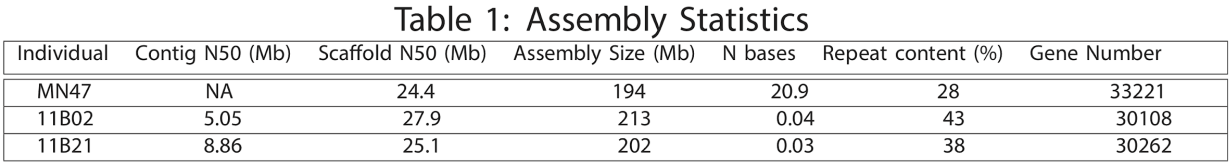
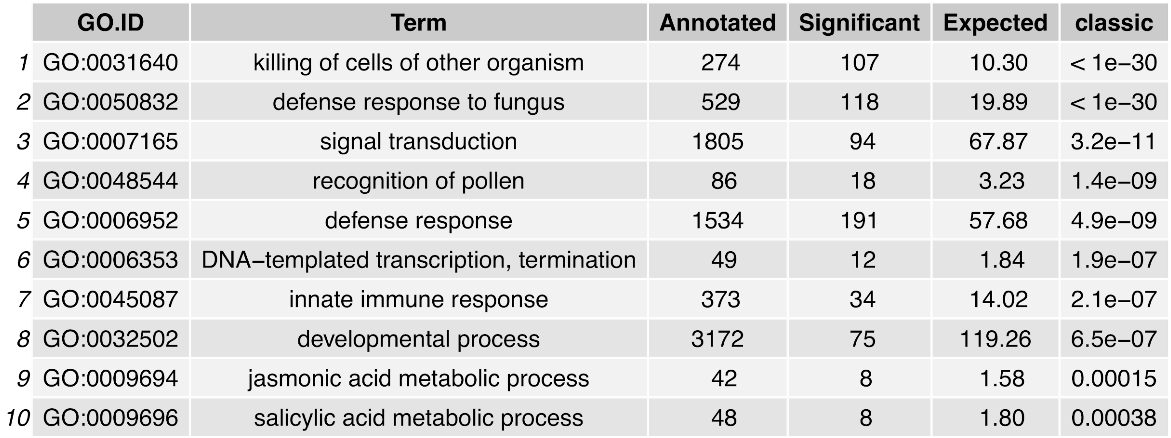


Table 1: Assembly Statistics

Individual	Contig N50 (Mb)	Scaffold N50 (Mb)	Assembly Size (Mb)	N bases	Repeat content (%)	Gene Number
MN47	NA	24.4	194	20.9	28	33221
11B02	5.05	27.9	213	0.04	43	30108
11B21	8.86	25.1	202	0.03	38	30262

Supplementary Table 1. Assembly statistics comparing the reference *A. lyrata* genomes and the two newly assembled *A. lyrata* genomes.



	GO.ID	Term	Annotated	Significant	Expected	classic
1	GO:0031640	killing of cells of other organism	274	107	10.30	< 1e−30
2	GO:0050832	defense response to fungus	529	118	19.89	< 1e−30
3	GO:0007165	signal transduction	1805	94	67.87	3.2e−11
4	GO:0048544	recognition of pollen	86	18	3.23	1.4e−09
5	GO:0006952	defense response	1534	191	57.68	4.9e−09
6	GO:0006353	DNA−templated transcription, termination	49	12	1.84	1.9e−07
7	GO:0045087	innate immune response	373	34	14.02	2.1e−07
8	GO:0032502	developmental process	3172	75	119.26	6.5e−07
9	GO:0009694	jasmonic acid metabolic process	42	8	1.58	0.00015
10	GO:0009696	salicylic acid metabolic process	48	8	1.80	0.00038

Supplementary Table 2. Genes missing from the annotation of the 11B genomes but present in *A. lyrata* reference and *A. thaliana* reference, likely reflect the limited RNAseq and conditions used to annotate genes in the 11B genomes, therefore a liftover of these genes was performed using the *A. lyrata* reference ortholog.

References

1. Greilhuber, J. *et al.* Smallest angiosperm genomes found in lentibulariaceae, with chromosomes of bacterial size. *Plant Biol.* **8**, 770–777 (2006).
2. Pellicer, J., Fay, M. F. & Leitch, I. J. The largest eukaryotic genome of them all? *Bot. J. Linn. Soc.* **164**, 10–15 (2010).
3. Meyer, A. *et al.* Giant lungfish genome elucidates the conquest of land by vertebrates. *Nature* (2021) doi:10.1038/s41586-021-03198-8.
4. Blommaert, J., Riss, S., Hecox-Lea, B., Mark Welch, D. B. & Stelzer, C. P. Small, but surprisingly repetitive genomes: transposon expansion and not polyploidy has driven a doubling in genome size in a metazoan species complex. *BMC Genomics* **20**, 466 (2019).
5. Soltis, D. E., Soltis, P. S., Bennett, M. D. & Leitch, I. J. Evolution of genome size in the

angiosperms. *Am. J. Bot.* **90**, 1596–1603 (2003).

6. Kidwell, M. G. Transposable elements and the evolution of genome size in eukaryotes. *Genetica* **115**, 49–63 (2002).

7. Sun, C. *et al.* LTR retrotransposons contribute to genomic gigantism in plethodontid salamanders. *Genome Biol. Evol.* **4**, 168–183 (2012).

8. Naville, M. *et al.* Massive Changes of Genome Size Driven by Expansions of Non-autonomous Transposable Elements. *Curr. Biol.* **29**, 1161–1168.e6 (2019).

9. Long, Q. *et al.* Massive genomic variation and strong selection in Arabidopsis thaliana lines from Sweden. *Nat. Genet.* **45**, 884–890 (2013).

10. Pellicer, J., Kelly, L. J., Leitch, I. J., Zomlefer, W. B. & Fay, M. F. A universe of dwarfs and giants: genome size and chromosome evolution in the monocot family Melanthiaceae. *New Phytol.* **201**, 1484–1497 (2014).

11. Shirasu, K., Schulman, A. H., Lahaye, T. & Schulze-Lefert, P. A contiguous 66-kb barley DNA sequence provides evidence for reversible genome expansion. *Genome research* vol. 10 908–915 (2000).

12. Devos, K. M., Brown, J. K. M. & Bennetzen, J. L. Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis. *Genome Res.* **12**, 1075–1079 (2002).

13. Nilsson, A. I. *et al.* Bacterial genome size reduction by experimental evolution. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 12112–12116 (2005).

14. Vu, G. T. H., Cao, H. X., Reiss, B. & Schubert, I. Deletion-bias in DNA double-strand break repair differentially contributes to plant genome shrinkage. *New Phytol.* **214**, 1712–1721 (2017).

15. Petrov, D. A. DNA loss and evolution of genome size in Drosophila. *Genetica* **115**, 81–91 (2002).

16. Kapusta, A., Suh, A. & Feschotte, C. Dynamics of genome size evolution in birds and mammals. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E1460–E1469 (2017).

17. Petrov, D. A. & Hartl, D. L. Pseudogene evolution and natural selection for a compact genome. *J. Hered.* **91**, 221–227 (2000).

18. Kuo, C.-H., Moran, N. A. & Ochman, H. The consequences of genetic drift for bacterial genome complexity. *Genome Res.* **19**, 1450–1454 (2009).

19. Bilinski, P. *et al.* Parallel altitudinal clines reveal trends in adaptive evolution of genome size in Zea mays. *PLoS Genet.* **14**, e1007162 (2018).

20. Beaulieu, J. M., Leitch, I. J., Patel, S., Pendharkar, A. & Knight, C. A. Genome size is a strong predictor of cell size and stomatal density in angiosperms. *New Phytol.* **179**, 975–986 (2008).

21. Wright, N. A., Gregory, T. R. & Witt, C. C. Metabolic 'engines' of flight drive genome size reduction in birds. *Proceedings of the Royal Society B: Biological Sciences* **281**, 20132780 (2014).

22. Sabath, N., Ferrada, E., Barve, A. & Wagner, A. Growth temperature and genome size in bacteria are negatively correlated, suggesting genomic streamlining during thermal adaptation. *Genome Biol. Evol.* **5**, 966–977 (2013).

23. Hu, T. T., Pattyn, P., Bakker, E. G., Cao, J. & Cheng, J. F. The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nature* (2011).

24. Fierst, J. L. *et al.* Reproductive Mode and the Evolution of Genome Size and Structure in Caenorhabditis Nematodes. *PLoS Genet.* **11**, e1005323 (2015).

25. Wright, S. I., Ness, R. W., Foxe, J. P. & Barrett, S. C. H. Genomic Consequences of Outcrossing and Selfing in Plants. *Int. J. Plant Sci.* **169**, 105–118 (2008).

26. Charlesworth, B., Charlesworth, D. & Morgan, M. T. Genetic loads and estimates of mutation rates in highly inbred plant populations. *Nature* **347**, 380–382 (1990).

27. Roessler, K. *et al.* The genome-wide dynamics of purging during selfing in maize. *Nat Plants* **5**, 980–990 (2019).

28. Nordborg, M. Linkage disequilibrium, gene trees and selfing: an ancestral recombination graph with partial self-fertilization. *Genetics* **154**, 923–929 (2000).

29. Boutin, T. S., Le Rouzic, A. & Capy, P. How does selfing affect the dynamics of selfish transposable elements? *Mob. DNA* **3**, 1–9 (2012).

30. Hollister, J. D. *et al.* Transposable elements and small RNAs contribute to gene expression divergence between Arabidopsis thaliana and Arabidopsis lyrata. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 2322–2327 (2011).

31. Charlesworth, D. & Charlesworth, B. Transposable elements in inbreeding and outbreeding populations. *Genetics* vol. 140 415–417 (1995).

32. Jiao, W.-B. & Schneeberger, K. Chromosome-level assemblies of multiple Arabidopsis genomes

reveal hotspots of rearrangements with altered evolutionary dynamics. *Nat. Commun.* **11**, 989 (2020).

33. Briskine, R. V. *et al.* Genome assembly and annotation of Arabidopsis halleri, a model for heavy metal hyperaccumulation and evolutionary ecology. *Mol. Ecol. Resour.* **17**, 1025–1036 (2017).

34. Burns, R. *et al.* Gradual evolution of allopolyploidy in Arabidopsis suecica. doi:10.1101/2020.08.24.264432.

35. Novikova, P. Y. *et al.* Sequencing of the genus Arabidopsis identifies a complex history of nonbifurcating speciation and abundant trans-specific polymorphism. *Nat. Genet.* **48**, 1077–1082 (2016).

36. de la Chaux, N., Tsuchimatsu, T., Shimizu, K. K. & Wagner, A. The predominantly selfing plant Arabidopsis thaliana experienced a recent reduction in transposable element abundance compared to its outcrossing relative Arabidopsis lyrata. *Mob. DNA* **3**, 2 (2012).

37. Zhang, S.-J., Liu, L., Yang, R. & Wang, X. Genome Size Evolution Mediated by Gypsy Retrotransposons in Brassicaceae. *Genomics Proteomics Bioinformatics* **18**, 321–332 (2020).

38. Rawat, V. *et al.* Improving the Annotation of Arabidopsis lyrata Using RNA-Seq Data. *PLoS One* **10**, e0137391 (2015).

39. Slotte, T. *et al.* The Capsella rubella genome and the genomic consequences of rapid mating system evolution. *Nat. Genet.* **45**, 831–835 (2013).

40. Nattestad, M. & Schatz, M. C. Assemblytics: a web analytics tool for the detection of variants from an assembly. *Bioinformatics* **32**, 3021–3023 (2016).

41. Nordborg, M. *et al.* The pattern of polymorphism in Arabidopsis thaliana. *PLoS Biol.* **3**, e196 (2005).

42. Quadrana, L. *et al.* The Arabidopsis thaliana mobilome and its impact at the species level. *Elife* **5**, (2016).

43. Lee, Y. C. G. & Karpen, G. H. Pervasive epigenetic effects of euchromatic transposable elements impact their evolution. *Elife* **6**, (2017).

44. Fawcett, J. A., Rouzé, P. & Van de Peer, Y. Higher Intron Loss Rate in Arabidopsis thaliana Than A. lyrata Is Consistent with Stronger Selection for a Smaller Genome. *Mol. Biol. Evol.* **29**, 849–859 (2011).

45. Yang, Y.-F., Zhu, T. & Niu, D.-K. Association of Intron Loss with High Mutation Rate in

Arabidopsis: Implications for Genome Size Evolution. *Genome Biol. Evol.* **5**, 723–733 (2013).

46. Rose, A. B. Introns as Gene Regulators: A Brick on the Accelerator. *Front. Genet.* **9**, 672 (2018).
47. He, F. *et al.* Cis-regulatory evolution spotlights species differences in the adaptive potential of gene expression plasticity. *Cold Spring Harbor Laboratory* 2020.07.02.184036 (2020) doi:10.1101/2020.07.02.184036.
48. Farlow, A., Meduri, E. & Schlötterer, C. DNA double-strand break repair and the evolution of intron density. *Trends Genet.* **27**, 1–6 (2011).
49. Roy, S. W. The origin of recent introns: transposons? *Genome Biol.* **5**, 251 (2004).
50. Vu, G. T. H. *et al.* Comparative Genome Analysis Reveals Divergent Genome Size Evolution in a Carnivorous Plant Genus. *Plant Genome* **8**, eplantgenome2015.04.0021 (2015).
51. Schwer, B. *et al.* Transcription-associated processes cause DNA double-strand breaks and translocations in neural stem/progenitor cells. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 2258–2263 (2016).
52. Griffin, P. C. & Willi, Y. Evolutionary shifts to self-fertilisation restricted to geographic range margins in North American Arabidopsis lyrata. *Ecol. Lett.* **17**, 484–490 (2014).
53. Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
54. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).
55. Sedlazeck, F. J. *et al.* Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **15**, 461–468 (2018).
56. Walker, B. J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963 (2014).
57. Alonge, M. *et al.* RaGOO: fast and accurate reference-guided scaffolding of draft genomes. *Genome Biol.* **20**, 224 (2019).
58. Sun, H., Ding, J., Piednoël, M. & Schneeberger, K. findGSE: estimating genome size variation within human and Arabidopsis using k-mer frequencies. *Bioinformatics* **34**, 550–557 (2018).
59. Flynn, J. M. *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 9451–9457 (2020).
60. SMIT & A., F. A. Repeat-Masker Open-3.0. *http://www.repeatmasker.org* (2004).

61. Bailly-Bechet, M., Haudry, A. & Lerat, E. 'One code to find them all': a perl tool to conveniently parse RepeatMasker output files. *Mob. DNA* **5**, 13 (2014).

62. Stanke, M. *et al.* AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–9 (2006).

63. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).

64. Gremme, G. GenomeThreader Gene Prediction Software. (2014).

65. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

66. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).

67. Haas, B., Papanicolaou, A. & Others. TransDecoder (find coding regions within transcripts). *Google Scholar* (2016).

68. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**, 157 (2015).

69. Alexa, A. & Rahnenfuhrer, J. topGO: enrichment analysis for gene ontology. *R package version* **2**, 2010 (2010).

70. Shumate, A. & Salzberg, S. L. Liftoff: accurate mapping of gene annotations. *Bioinformatics* (2020) doi:10.1093/bioinformatics/btaa1016.

71. Marçais, G. *et al.* MUMmer4: A fast and versatile genome alignment system. *PLoS Comput. Biol.* **14**, e1005944 (2018).

Discussion

This thesis has focused on a tale of two transitions: polyploidization and changes in mating type, both of which have often been referred to in the literature as “evolutionary dead ends”[1,2]. Yet despite the theoretical predictions, many polyploid and self-compatible (and often a combination of the two) plant species exist and are evolutionarily successful. We study genomes from the *Arabidopsis* genus to explore the genetic consequences of these evolutionary transitions.

Natural polyploids likely experience a selective bottleneck

We studied the process of polyploidization in the natural allotetraploid species *A. suecica*, generated ~16kya through the hybridization of *A. thaliana* and *A. arenosa*. We find that rather than a single origin, multiple individuals contributed to the gene pool of *A. suecica*. The multiple origins of *A. suecica* is likely a key reason why the species was able to become established and survive the extreme bottleneck associated with its origin, a striking example of which is the ~8Mb on chromosome 2 of the *A. suecica* genome that is devoid of variation. The evolution of self-compatibility likely allowed *A. suecica* to also avoid extinction by guaranteeing reproductive success despite having a minority cytotype at the time of its origin. Through generating a chromosome-level genome sequence and transcriptome data for multiple individuals we also do not find evidence for any genome shock in *A. suecica*: the genome is highly collinear, TEs are not mis-regulated and there is no subgenome dominance in gene expression. On the contrary, we find evidence for adaptation towards a "stable" polyploid life, with genes involved in cell division being up-regulated on the *A. thaliana* subgenome and plastid related genes up-regulated on the *A. arenosa* subgenome. Given that synthetic *A. suecica* do show signs of genome rearrangements and are often aneuploid we suggest that the biggest bottleneck in new polyploid species is likely selective. Thus, the domesticated species that show signatures of genome shock in the form of large genome rearrangements and subgenome dominance may in fact be “hopeful monsters" that have persisted because of humans. An argument against this hypothesis is the existence of natural polyploids such as like *Tragopogon* and *Mimmulus* which do show signs of genome shock in nature, however, we argue that as these new polyploids are very young (less than 200 years old) such that they may in fact currently be experiencing a selective bottleneck.

Many questions remain how the subgenomes of *A. suecica* despite being separated by over 6 million years and their difference in genome size can function together. The observation that the plastid genes from the *A. arenosa* subgenome are up-regulated rather than silenced raises questions of how these nuclear encoded genes are interacting with the maternally inherited chloroplasts from *A. thaliana*. Sequencing of chloroplast transcripts from *A. suecica* could shed more light on the types of organelle genes that are involved as well as the generation of CRISPR mutants to determine any phenotypic effects of their knockout. An ideal experiment would involve switching the direction of the cross that produces *A. suecica*. While the mechanism that restricts the direction of the cross is largely unknown, the phenomena is known as the ``SIxSC rule"[3], in which SI (self incompatible) pollen tends to be able to fertilize SC (self compatible) pistils but not vice versa. Studies in selfing and outcrossing tomatoes have linked the phenomena to a gene (Cullin1) that is expressed in pollen[4] and the gene notably interacts with a gene at or near the S-locus. Identifying similar genes in *A. suecica* could allow for bi-directional crosses to be obtained.

The observation that a rare homeologous exchange event can delete 100 genes without causing an obvious phenotypic change raises the question of how much of the genome is essential and whether such events play a key role in the re-diploidization process in allopolyploids. Sequencing of additional *A. suecica* genomes would help answer the variation in HE, the rate at which it occurs and the genetic factors that may be involved.

Genome size evolution and selfing

We studied the process of genome shrinkage by comparing multiple genomes of different *Arabidopsis* species that include both mating types: selfing and outcrossing. By using multiple individuals and species we aimed to determine the evolutionary history of the length variants involved in the reduced genome size of the predominantly selfing *A. thaliana* compared to the larger genomes of its outcrossing relatives.

We find no evidence for selection on standing variation playing a role in genome size reduction in *A. thaliana*, suggesting that the mutations involved are likely *de novo*. The shrinkage of the genome and chromosomal rearrangements likely occurred after the species transitioned to selfing as theory predicts that rearrangements that are otherwise deleterious in outcrossers can fix at a faster rate in highly selfing species[5], though it remains unknown if the karyotype change and genome shrinkage occurred multiple times independently or occurred once in *A. thaliana*. Estimates of the transition to sefing in *A. thaliana* are ~1 Mya[6], which could suggest that the genome size reduction and change in chromosome number are recent events. Notably the genome rearrangements also act as a strong reproductive barrier since F1 hybrids between *A. thaliana* and the outcrossing species are sterile[7]. The recent transition to selfing could also explain the multiple negative length variants we find still segregating in *A. thaliana*, as selection for a compact genome may still be ongoing.

However, the fact that multiple alleles also exist in *A. lyrata* and the outcrossing species likely indicates that these regions were also highly variable in the ancestor. Therefore we suggest that these genomic regions are unstable regions. A scenario of mutation-selection balance may be involved, in which mutations occurring at these genome regions are frequent and that negative length variants are being selected for in *A. thaliana*, while in the outcrossers the same level of selection is not taking place with both positive and negative length variants segregating in these genomes. As we could not infer the correct ancestral state of the length variants due to their highly polymorphic nature (i.e. we could not conclusively infer whether the negative length variants are a derived deletion in *A. thaliana* or that the positive length variants are a derived insertion in *A. lyrata*) it remains difficult to conclude how a genome shrinks. Analysis of more complete genomes from many *A. thaliana* individuals may be a way forward to infer the ancestral state. By examining hundreds of genomes, the majority allele can be inferred as ancestral and the minor allele can be considered derived. On a local level one can examine how positive and negative alleles segregate locally in subpopulations to help answer if the *A. thaliana* genome is still shrinking.

Determining why these sites may be prone to mutations or double strand breaks, an experiment in which genomic regions that are highly polymorphic in length in *Arabidopsis* could be inserted into yeast artificial chromosomes upstream of a reporter gene, as was done for investigating fragile DNA in sticklebacks[8]. Sequencing of transgenic loci after targeted DSBs in the different species of *Arabidopsis* could be carried out to determine if a deletion bias of DSBs exists in *A. thaliana* compared to its outcrossing *Arabidopsis* relatives, as was indicated in a comparison between *A. thaliana* and barley[9].

A. suecica also provides a unique opportunity in studying the question of genome shrinkage, as the *A. arenosa* subgenome is originally from an outcrossing ancestor and has been in the genome of a selfer for ~16kya. Analyzing the evolutionary history of the *A. arenosa* subgenome is first required in order to compare *A. suecica* with the most closely related *A. arenosa* population, though we note here that TEs on the *A. arenosa* subgenome of *A. suecica* are fewer in number, consistent with expectations for a selfer. Following synthetic *A. suecica* lines and examining changes in the *A. arenosa* subgenome over multiple generations could also help answer this question by providing the immediate genome effects of selfing. Finally, linking changes seen over the short term in *A. suecica* with the changes over the long term between *A. thaliana* and *A. lyrata* could provide mechanistic insight into how genomes shrink and how repeatable genome size reduction is.

1. Igic, B. & Busch, J. W. Is self-fertilization an evolutionary dead end? *New Phytol.* **198**, 386–397 (2013).
2. Soltis, D. E. *et al.* Are polyploids really evolutionary dead-ends (again)? A critical reappraisal of Mayrose et al.(2011). *New Phytol.* **202**, 1105–1117 (2014).
3. Brandvain, Y. & Haig, D. Divergent mating systems and parental conflict as a barrier to hybridization in flowering plants. *Am. Nat.* **166**, 330–338 (2005).
4. Li, W. & Chetelat, R. T. A Pollen Factor Linking Inter- and Intraspecific Pollen Rejection in Tomato. *Science* vol. 330 1827–1830 (2010).
5. Charlesworth, B. Evolutionary rates in partially self-fertilizing species. *Am. Nat.* **140**, 126–148 (1992).
6. Durvasula, A. *et al.* African genomes illuminate the early history and transition to selfing in Arabidopsis thaliana. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 5213–5218 (2017).
7. Nasrallah, M. E., Yogeeswaran, K., Snyder, S. & Nasrallah, J. B. Arabidopsis species hybrids in the study of species differences and evolution of amphiploidy in plants. *Plant Physiol.* **124**, 1605–1614 (2000).
8. Xie, K. T. *et al.* DNA fragility in the parallel evolution of pelvic reduction in stickleback fish. *Science* **363**, 81–84 (2019).
9. Vu, G. T. H., Cao, H. X., Reiss, B. & Schubert, I. Deletion-bias in DNA double-strand break repair differentially contributes to plant genome shrinkage. *New Phytol.* **214**, 1712–1721 (2017).