

On the genetics and genomics of *Arabidopsis thaliana* and its relatives

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

Danelle Kathleen Seymour

aus Oakland, California, USA

Tübingen

2016

Tag der mündlichen Qualifikation:

29.04.2016

Dekan:

Prof. Dr. Wolfgang Rosenstiel

1. Berichterstatter

Prof. Dr. Detlef Weigel

2. Berichterstatter

Prof. Dr. Marja Timmermans

Table of Contents

SUMMARY OF DISSERTATION RESEARCH	7
ZUSAMMENFASSUNG DER DISSERTATIONSARBEIT	9
PUBLICATIONS.....	11
1. INTRODUCTION	12
MUTATIONAL PROCESSES LEAD TO VARIATION IN GENOME ARCHITECTURE.....	12
SELECTION SHAPES GENOME ARCHITECTURE.....	14
THE CONTRIBUTION OF GENETIC DRIFT TO GENOME ARCHITECTURE.....	16
SPECIATION: A CONSEQUENCE OF MUTATION, SELECTION, AND GENETIC DRIFT	17
2. OBJECTIVES OF DOCTORAL RESEARCH	18
I. INTERSPECIFIC TE PROLIFERATION AND COMPARATIVE EPIGENOMICS.....	18
II. THE CONTRIBUTION OF MUTATION-SELECTION BALANCE TO HYBRID PHENOTYPES	19
III. INTRASPECIFIC ALLELIC DISTORTION AND THE MOLECULAR ARMS RACE.....	20
3. “EVOLUTION OF DNA METHYLATION PATTERNS IN THE BRASSICACEAE IS DRIVEN BY DIFFERENCES IN GENOME ORGANIZATION”	22
4. “IMPROVING THE ANNOTATION OF <i>ARABIDOPSIS LYRATA</i> USING RNA-SEQ DATA”	82
5. “THE GENETIC ARCHITECTURE OF NON-ADDITIVE HYBRID PHENOTYPES IN <i>ARABIDOPSIS THALIANA</i> HYBRIDS”	96
6. “RECURRENT SEGREGATION DISTORTION IS UNCOVERED IN A SPECIES-WIDE SCREEN FOR BIASED GENETIC TRANSMISSION”	246
7. DISCUSSION	285
I. THE EVOLUTION OF DNA METHYLATION IN THE BRASSICACEAE	285
II. THE GENETIC ARCHITECTURE OF NON-ADDITIVE HYBRID PHENOTYPES	289
III. A SPECIES-WIDE SCREEN FOR INTRASPECIFIC GENETIC BARRIERS	291
8. REFERENCES	294
9. ACKNOWLEDGEMENTS	300

Summary of dissertation research

During my dissertation, I pursued a three-pronged approach to characterize the consequences of plant genome evolution. Genomes are sculpted by the combined influences of mutation, selection, and genetic drift. As a result of these processes, genome size, as well as the overall architecture of genomes, is constantly fluctuating. A species' genomic architecture can impact what types of genetic variants give rise to phenotypic variation and may even influence the ability of a species to adapt to new environments.

Transposable elements (TEs) are a major source of variation in genome architecture and the fast evolution of these elements can swiftly alter a species' epigenetic landscape as a result of their tightly linked epigenetic marks, including DNA methylation. The first aim was to characterize the contribution of DNA methylation to variation in transcriptional levels, a molecular phenotype that is perturbed by the close proximity of methylated TEs. Using a comparative genomics approach, I found that interspecific variation in DNA methylation is driven by the lineage-specific evolution of underlying TE sequences and that, as a result, any transcriptional consequences do not persist over the long term.

At any particular point in time, a species will contain a suite of deleterious mutations in the process of elimination from the population and mutation-selection balance will dictate the rate at which this occurs. Whole-genome complementation of these mutations by outbreeding is hypothesized to give rise to hybrid vigor, also known as heterosis. Using a large collection of first generation hybrids, I characterized the genetic architecture of vigorous hybrid phenotypes to estimate the contribution of deleterious mutations to phenotypic variation. I found that processes other than mutation-selection balance, either genetic bottlenecks or adaptive processes, also give rise to the variants underlying hybrid phenotypes.

Finally, I sought to understand the species-wide frequency and underlying genetic basis of intraspecific genetic barriers that arise as a byproduct of genome evolution. The relative contribution of stochastic versus adaptive processes to hybrid dysfunction, an irreversible step towards

speciation, is unknown. I surveyed the species-wide rate of segregation distortion, a molecular signature of hybrid dysfunction, using a large set of genetically diverse F_2 populations. Distorted loci were uncovered in 12-24% of segregating populations, an indication that the biased transmission of alleles is not a one-off genetic anomaly, but may contribute substantially to the formation of genetic barriers.

Zusammenfassung der Dissertationsarbeit

Die vorliegende Dissertationsarbeit hatte als Ziel die Charakterisierung der genomischen und phänotypischen Auswirkungen der pflanzlichen Genomevolution und ist unterteilt in drei getrennte, ineinander greifende Ansätze. Das Pflanzengenom unterliegt den Einflüssen von Mutation, Selektion und Gendrift. Dadurch sind die Größe und Gesamtarchitektur des Genoms in konstantem Fluss. Die Genomarchitektur einer Spezies kann bestimmend dafür sein, welche genetischen Varianten zu einer phänotypischen Ausprägung führen; sie kann sogar die Fähigkeit der Spezies beeinflussen, sich an veränderte Umweltbedingungen anzupassen.

Eine hauptsächliche Quelle für Variationen in der Genomarchitektur sind Transposons, mobile genetische Elemente, die auch als „springende Gene“ bezeichnet werden. Transposons sind mit epigenetischen Markern verknüpft, daher kann die schnelle Evolution dieser genetischen Elemente auch zu beschleunigter Veränderung der epigenetischen Umgebung führen, inklusive Veränderungen der DNA Methylierung. Dies wiederum kann einen molekularen Phänotyp nach sich ziehen, indem die Transposon-Methylierung einen Effekt auf die Expression von Genen in direkter Nachbarschaft zum Transposon ausübt. Das erste Ziel der vorliegenden Arbeit war es daher, den Einfluss der DNA Methylierung auf die Variation der Transkriptmenge zu charakterisieren. Ein vergleichender genomischer Ansatz ergab, dass interspezifische Variation der DNA Methylierung ihren Ursprung in der artspezifischen Evolution von Transposonsequenzen hat, und somit transposonbedingte transkriptionelle Effekte nicht über Arten hinweg bestehen bleiben.

Eine Spezies beinhaltet zu jedem Zeitpunkt eine Reihe von schädlichen („*deleterious*“) Mutationen, die dabei sind, aus der Population zu verschwinden. Die Rate, mit der dies geschieht, wird bestimmt durch das sogenannte Mutations-Selektions-Gleichgewicht („*mutation-selection balance*“), bei dem die Entstehung neuer und das Verschwinden bestehender schädlicher Mutationen sich in ihrer Häufigkeit aufheben. Es wird vermutet, dass Heterosis-Effekte aus der Komplementierung dieser Mutationen durch Auskreuzen entstehen. In der vorliegenden Arbeit wurde, unter Zuhilfenahme

einer ausgedehnten Sammlung an Hybriden der ersten Generation (F_1), der Beitrag schädlicher Mutationen zur phänotypischen Variation untersucht. Es wurde fest gestellt, dass andere Prozesse als das Mutations-Selektions-Gleichgewicht – entweder ein genetischer Flaschenhals („*genetic bottleneck*“) oder adaptive Prozesse – zur Entstehung von Varianten beitragen, die dem Heterosis-Effekt zugrunde liegen.

Als abschließenden Aspekt untersuchte die vorliegende Arbeit die genetischen Grundlagen und die Häufigkeit intraspezifischer Barrieren. Diese können als Nebenprodukt der Genomevolution auftreten. Die Dysfunktion von Hybriden ist ein irreversibler Schritt während der Artbildung; der relative Beitrag stochastischer *versus* adaptiver Prozesse hierbei ist jedoch unbekannt. In meiner Dissertation untersuchte ich die spezies-weite Rate von Genomkonflikten („*segregation distortions*“) anhand genetisch unterschiedlicher F_2 Populationen. Solche Genomkonflikte sind eine molekulare Signatur für Hybrid-Dysfunktionen. In 12-24% der segregierenden Populationen konnten entsprechende genomische Loci gefunden werden, ein Hinweis darauf, dass die ungleiche Weitergabe der Allele keine seltene genetische Anomalie darstellt, sondern im Gegenteil substantiell zur Entstehung genetischer Barrieren beiträgt.

Publications

Accepted papers

Seymour DK*, Koenig D*, Hagmann J, Becker C, Weigel D (2014). Evolution of DNA methylation patterns in the Brassicaceae is driven by differences in genome organization. *PLoS Genet* 10:e1004785.

Rawat V, Abdelsamad A, Pietzenuk B, **Seymour DK**, Koenig D, Weigel D, Pecinka A, Schneeberger K (2015). Improving the annotation of *Arabidopsis lyrata* using RNA-seq data. *PLoS One* 10:e0137391.

Submitted papers

Seymour DK*, Chae E*, Grimm DG*, Martín-Pizarro C, Vasseur F, Rakitsch B, Borgwardt K, Koenig D, Weigel D. The genetic architecture of non-additive hybrid phenotypes in *Arabidopsis thaliana*. (In revision).

Rowan BA, **Seymour DK**, Chae E, Lundberg D, Weigel D. Methods for genotyping-by-sequencing. *Methods in Molecular Biology* (In press).

Ready to submit manuscripts

Seymour DK, Chae E, Ariöz B, Koenig D, Weigel D. Recurrent segregation distortion is uncovered in a species-wide screen for biased genetic transmission.

Additional publications not included in this dissertation

Karlsson P, Christie MD, **Seymour DK**, Wang H, Wang X, Hagmann J, Kulcheski F, Manavella P (2015). KH domain protein RCF3 is a tissue-biased regulator of the plant miRNA biogenesis cofactor HYL1. *Proc Natl Acad Sci USA* 112:14096-14101.

Rugnone ML, Faigón Soverna A, Sanchez SE, Schlaen RG, Hernando CE, **Seymour DK**, Mancini E, Chernomoretz A, Weigel D, Más P, Yanovsky MJ (2013). LNK genes integrate light and clock signaling networks at the core of the *Arabidopsis* oscillator. *Proc Natl Acad Sci USA* 110:12120-12125.

*These authors contributed equally to this work

1. Introduction

A species' genomic architecture is shaped by the sum of multiple evolutionary processes, including mutation, selection, and genetic drift. Genome architecture, or the placement of genetic elements onto a genome, refers to a number of characteristics, including genome size, karyotype, and genome organization, for example, gene order and gene density (reviewed in [1]). Plant genomes are particularly pliable. Although gene number is relatively stable across millions of years of evolution, the genome size and chromosome number of plant species ranges over 2,000-fold and 300-fold, respectively [2-4]. This is in contrast to mammalian genomes, whose size only varies 5-fold [5]. This vast range of genome sizes is thought mostly to result from ancient polyploidy events, however these events are not sufficient to explain the magnitude of observed variability and other processes, such as mutation, recombination, and genetic drift, also contribute to rapid shifts in genome architecture (reviewed in [1, 6, 7]). In this introduction, I will discuss the forces that shape genome architecture, together with their associated consequences, and by doing so I hope to unite the three veins of research pursued during my dissertation work.

Mutational processes lead to variation in genome architecture

Whole genome duplication, or polyploidy, can quickly induce genome size variation, as these events increase an organism's DNA content by two-fold in only a single generation. Polyploidy is extremely common in plant species and 60-70% percent of modern species originate from polyploid ancestors [7-10] (reviewed in [11]). It is thought that such events allow organisms to adapt extremely rapidly, particularly in the face of a rapidly changing environment [12] (reviewed in [11, 13, 14]). Despite the frequency of polyploidy events and their adaptive potential, there are consequences associated with genome duplication, specifically meiotic dysfunction, that can lead to inviability and sterility [15, 16]. Typically, perhaps as a result of those costs, the polyploid genome set quickly erodes to functional diploidy. Mirroring that observation, there are only four ancient whole-genome duplication events

whose remnants are preserved in modern plant genomes [17-19]. Despite the obvious contribution of polyploidy to rapid genome size shifts, these whole-genome duplications quickly decay and additional genetic processes must be invoked to sufficiently explain the immense variation in genome architecture observed in plant genomes.

Mutational processes, including error-prone replication, aberrant recombination, and transposon proliferation, also contribute substantially to genome architecture variation. DNA polymerase slippage during replication can result in rapid expansion and contraction of simple satellite repeats (reviewed in [20]), and although its contribution to genome size expansion is relatively minor, repeat variation provides a template for secondary mutagenic events, primarily aberrant recombination (reviewed in [20]). Recombination is a critical evolutionary process whose benefits are two-fold. First, recombination ensures faithful segregation of genetic material into gametes and, second, it facilitates the shuffling of genetic material to prevent a species' evolutionary stagnation. All biological processes are imperfect, and recombination errors, including unequal crossing over, intrastrand crossing over, or non-homologous recombination, can lead to gene duplication or gene loss, repeat expansion, inversions, and translocations (reviewed in [1]). When double stranded breaks, the cellular signal to initiate recombination events, are not repaired, massive chromosomal changes can result. The cell's response to such events is to repair the break at any cost via the non-homologous end joining (NHEJ) pathway (reviewed in [20]). This pathway will join any free DNA ends and can result in chromosome fusion, breakage, or even loss. All types of recombination-driven genomic rearrangements occur frequently, particularly in complex genomic regions such as at repetitive or duplicated sequences (reviewed in [1]). In support of this, there are a number of large-scale rearrangements in the Brassicaceae, a family consisting of 3,600 member species [21] that vary substantially in both karyotype and genome size [22], and *in situ* hybridization experiments in multiple family members have shown that rearrangement breakpoints are enriched near repetitive sequences [23, 24].

Transposable elements (TEs) are likely the largest contributor to genome size variation [25]; not only do TEs proliferate rapidly, but they also

provide an enormous genomic target for illicit recombination. Transposons are insertional mutagens that selfishly propagate themselves via two major mechanisms (reviewed in [26]). Retrotransposons, or Class I elements, multiply via a “copy-and-paste” mechanism and, as a result, are able to rapidly proliferate. This class of transposons is the most common in plant genomes, comprising 7 to 75% of their total length [25]. Class II elements, consisting of DNA transposons, mobilize using a “cut-and-paste” mechanism. While these events unlikely lead to rapid expansion of repeated sequences, they are often mutagenic, as excision is imperfect, and they can also carry short stretches of flanking host DNA to their new location (reviewed in [26]). The final category of TEs, helitrons, replicate using rolling circle amplification and, similar to retrotransposons, their copy number increases with each transposition event (reviewed in [26]). Although these latter categories of TEs are not as successful as retrotransposons, they can comprise up to 5% of plant genomes [25], which in some cases is equivalent to the protein coding sequence content [25].

Selection shapes genome architecture

Not only is recombination mutagenic, but recombination rates also mediate the effects of selection and together they shape the evolutionary fate of new mutations. As a result, genomes should be organized along recombinational gradients, with highly recombining regions enriched in functional elements (reviewed in [1]). This correlation seems to hold true in a number of genomes with a few exceptions, notably in maize telomeres where recombination is high but gene density is very low [27]. One expectation of this observation is that non-recombining regions should be enriched for mutagenic events, as selection cannot effectively act in these regions [28-30] (reviewed in [1]). This phenomenon, known as Muller’s ratchet, typically describes asexual organisms, where, in the absence of recombination, deleterious mutations cannot be purged from the population [30]. However, it also holds true in non-recombining regions of sexual species, as these regions are effectively asexual. The enrichment of mutations in non-

recombining regions has been demonstrated for a number of species [31-34], and the accumulation of such sites has important implications for fitness.

Selection, via recombination, can drastically alter genome organization, in particular, it has a profound effect on the distribution of sequences found in eu- and heterochromatin (reviewed in [1]). In plant genomes, the majority of TEs are localized in heterochromatic sequences, mainly centromeric and pericentromeric regions [35-39]. While this could be the result of biased insertion [40-42], there is evidence that modern TE insertions are preferentially purged from euchromatin as they may have deleterious effects [32, 43-46]. In heterochromatic regions, not only are TEs less likely to be deleterious as there is a paucity of genes in these regions, but they cannot be effectively purged from the genome due to the lack of recombination (reviewed in [1]). Many genomes are organized in this fashion, with TEs enriched in non-recombining, gene sparse, heterochromatic regions, but there are exceptions. Recent bursts of transposition can alter this landscape [46], however, there is evidence that euchromatic enrichment of TEs will decay over time [32, 43-46]. This has been shown for long terminal repeat (LTR) retrotransposons, as the time of their insertion can be estimated from the accumulation of mutations in their paired terminal repeats [46, 47]. Older insertions are typically located in heterochromatic regions, while new insertions are also found in euchromatin, suggesting that new events are purged over time [48].

The mutagenic potential of TEs is not restricted to the disruption of genic sequences, but epigenetic marks associated with TEs can also have deleterious effects. Hosts deploy DNA methylation and other silencing epigenetic marks to mitigate TE proliferation (reviewed in [49, 50]). These marks can effectively silence transcription of TEs, but there are often unintended transcriptional consequences for flanking endogenous host genes. In many plant species, methylated TEs dampen the expression of neighboring genes and the potency of this effect decays with increased physical distance between the two elements [51, 52]. In support of their deleterious effect, TEs located near genes are often found at low frequencies in populations and signatures of purifying selection have been linked to these sequences [48].

The contribution of genetic drift to genome architecture

The impact of genetic drift on a population is dependent on its effective population size, N_e . N_e refers to the size of a population that would experience drift at the same rate as the observed population. In an ideal population, and in the absence of selection, once a mutation has reached intermediate frequency, it should have an equal probability of fixation or loss. Demographic factors, including population bottlenecks, as in humans, or shifts to inbreeding, as in many plant species, can increase the impact of genetic drift in populations, and new mutations may not be purged from the population as quickly as expected, especially if they are nearly neutral (reviewed in [53]).

Genetic drift may play an important role in shaping the variation in genome size and architecture observed between prokaryotes and multicellular eukaryotes [54-56] (reviewed in [57]). Prokaryotic and viral genomes are typically small in size and comprised primarily of coding and regulatory sequences [54, 58]. These genomes contain few introns and non-regulatory intergenic sequences. Because these species have enormous effective population sizes, usually greater than 10^7 , any non-essential sequences are quickly purged from the genome, as the maintenance of such sequences has an energetic cost [54-56] (reviewed in [57]). This will continue as long as the strength of purifying selection acting on these sequences is less than $1/N_e$ (reviewed in [57]). As one moves into eukaryotes, N_e decreases which concomitantly increases the effect of genetic drift, enabling sequences with slightly deleterious effects to be maintained in the genome [54, 55]. Although eukaryotic genomes vary in their non-essential sequence content, there is certainly a trend for an increased number of introns per gene, and longer intergenic regions as N_e decreases [54, 55] (reviewed in [57]). Unicellular eukaryotes with large-effective population sizes typically have only a single intron, while humans have an average of seven introns per gene and a much smaller N_e [59]. The correlation of effective population size with exon-intron structure and non-essential sequence content is intriguing and it is likely that genetic drift played a role in shaping these genomic features.

Not only does genetic drift contribute to genome organization, but the severity of drift can have important consequences for the mode of selection

experienced by a population (reviewed in [60]). Evolutionary biologists have long been interested in the relative contribution of various modes of selection to phenotypic variation as well as their respective genomic footprints. There is particular interest in dissecting phenotypic variation into that caused by genetic drift, or mutation-selection balance, versus adaptive evolution, either directional or balancing selection. While there are empirical examples in support of both cases, the general influence of each is still unknown (reviewed in [60]). However, as more genomic data becomes available, it is apparent that demography heavily influences the contribution of each factor to phenotype. Species with large effective populations sizes, such as *Drosophila melanogaster*, evolve primarily under adaptive forces. In support of this, 40-90% of amino acid substitutions in this species are estimated to be adaptive [61, 62]. In contrast, phenotypic variation in small, or inbreeding, populations is driven predominantly by genetic drift as weakly deleterious mutations are permitted to accumulate in these scenarios. In humans, which have experienced a recent population bottleneck, only 10% of amino acid substitutions are predicted to be adaptive and a number of human diseases are the result of low frequency, harmful mutations [63, 64]. Genetic drift can dictate the evolutionary trajectory of a species, and its consequences are evident at many levels of genomic organization.

Speciation: a consequence of mutation, selection, and genetic drift

All of the evolutionary forces that I've discussed above - mutation, selection, and genetic drift - have been shown to cause hybrid dysfunction, an important step on the path toward speciation (reviewed in [65]). While speciation was traditionally thought to occur through adaptation to specific ecological niches [66], recent work has shown that the first steps in speciation may not be adaptive and there are a number of routes to hybrid dysfunction (reviewed in [65]). Mutational events, including gene duplication, followed by reciprocal loss of duplicates in independent lineages, can lead to the absence of either copy in subsequent progeny [67-69]. These events are facilitated by polyploidy, which can also lead to chromosomal fusion and translocations, another source of hybrid dysfunction [70, 71]. Relaxed selection and the

accumulation of mutations, due to mutation-selection balance, can also result in incompatible interactions and hybrid dysfunction [72-74]. Finally, selective forces can drive rapid evolution of particular sequences leading to functional divergence of interacting proteins. This can occur as a result of the host-pathogen arms race, where host proteins are subject to bouts of directional selection, or the molecular arms race occurring within genomes between selfish DNA and the host (reviewed in [65]). There are a number of examples of both. In plants, incompatible combinations of disease-resistance proteins, which identify and neutralize foreign, pathogen-injected proteins, result in hybrid dysfunction and these proteins display signatures of fast evolution [75-77]. In *Drosophila melanogaster*, a system where speciation is well studied, interspecific crossing perturbs the coevolution between selfish elements and their host which releases the element from its suppressed state and results in hybrid dysfunction and lethality [78, 79]. The evolutionary forces that give rise to variation in genomic architecture can have unintended consequences and may be the initial phase in the divergence between species.

2. Objectives of doctoral research

During the course of my dissertation research I've sought to understand the forces that shape genome architecture, together with their genomic and phenotypic consequences, in three independent research avenues.

1. Interspecific TE proliferation and comparative epigenomics

First, I used a comparative approach to examine the impact of TE proliferation on the (epi)-genomic landscape in three closely related Brassicaceae species. Evolutionary comparisons are especially insightful because they determine which biological phenomenon are important over time. Epigenetic marks, such as DNA methylation, are tightly linked to TEs and interest in the contribution of such marks to phenotypic variation has grown in recent years. TE-linked methylation can spread into nearby regions,

perturbing the expression of endogenous host genes [51, 52]. Despite empirical evidence for this phenomenon, the magnitude of the contribution of DNA methylation to gene expression variation, as well as the duration of such changes, is largely unknown. To examine the role of DNA methylation in generating long-term, stable phenotypic variation, I leveraged the power of comparative genomics. In all species, TEs are enriched in heterochromatic regions, but these sequences are so rapidly evolving that there is little conservation of TEs at the sequence level. Similarly, silencing epigenetic marks linked to TEs are also not conserved across lineages. One study species, *Arabidopsis lyrata*, has experienced a recent burst of TEs and invasion of modern TEs into euchromatic [46] has markedly altered the distribution of epigenomic marks in this species. In summary, rapid TE proliferation and turnover has shaped the genome architecture of these species and modern TE invasions are able to swiftly transform the epigenetic environment.

The genomic resources available in non-model systems, including *Arabidopsis lyrata* and *Capsella rubella*, are often not on par with resource-rich model systems, such as *Arabidopsis thaliana*. As a complement to this comparative genomics study, the *A. lyrata* transcriptional data was leveraged to improve upon the first version of this species' annotation. The upgraded annotation increases the number of protein coding gene models, resolves chimeric or incomplete gene models, and also locates miRNA and tRNA encoding genes. Continual improvement of these genomic resources is necessary to ensure that any associated analyses are of high quality.

II. The contribution of mutation-selection balance to hybrid phenotypes

Next, I sought to understand the contribution of mutation-selection balance to intraspecific phenotypic variation using a large collection of *Arabidopsis thaliana* F₁ hybrids as a study system [77]. Inbreeding individuals taken from natural outbreeding populations for one or more generations will homozygose weakly deleterious, recessive mutations and give rise to inferior progeny, a phenomenon known as inbreeding depression (reviewed in [80]). Naturally inbreeding plant species, such as *A. thaliana*, have likely purged

many of these alleles during the shift to an alternative mating system, but this process is imperfect, especially in regions of low recombination (reviewed in [80]). The opposite of inbreeding depression, known as heterosis or hybrid vigor, occurs upon intercrossing and is thought to be due to complementation of deleterious alleles [81-83]. However, it has been argued that adaptive processes, such as heterozygote advantage, may be driving this phenomenon [84, 85]. To understand the relative contribution of each process to vigorous hybrid phenotypes, I performed a genome-wide association mapping study in a genetically diverse set of hybrids and identified loci with both classical dominant and overdominant effects on hybrid phenotypes. That loci with considerable effect sizes were mapped does not rule out the possibility that many undetected, recessive mutations also underlie hybrid phenotypes. However, it does indicate that either large-effect deleterious mutations survived the demographic bottleneck induced by the shift to inbreeding or that single loci displaying heterozygote advantage are involved, suggesting a role for adaptive processes in generating vigorous hybrid phenotypes.

III. Intraspecific allelic distortion and the molecular arms race

Both inter- and intraspecific hybrid dysfunction has been documented in a number of systems and such events are important for both the initiation and maintenance of reproductive isolation (reviewed in [65]). Previous work has shown that hybrid dysfunction can arise from a number of biological processes, many of which are the byproducts of genome evolution (reviewed in [65]). Despite these examples, a clear understanding of the frequency of hybrid dysfunction and the relative contribution of each process is lacking. One molecular signal of hybrid dysfunction is allelic distortion in segregating populations. I characterized the frequency of this bias, known as segregation distortion, in over 500 segregating F₂ populations of *Arabidopsis thaliana* in order to gain insight into the biological processes driving hybrid dysfunction. Segregation distortion is common in *A. thaliana*, with 12-24% of surveyed populations exhibiting significant distortion in at least one genomic region. While molecular characterization of causal loci is ongoing, this survey

provides the first species-wide characterization of segregating genetic barriers.

3. “Evolution of DNA methylation patterns in the Brassicaceae is driven by differences in genome organization”

Seymour DK*, Koenig D*, Hagemann J, Becker C, Weigel D (2014).

*These authors contributed equally to this work

PLoS Genetics 10:e1004785. doi: 10.1371/journal.pgen.1004785.

Abstract

DNA methylation is an ancient molecular modification found in most eukaryotes. In plants, DNA methylation is not only critical for transcriptionally silencing transposons, but can also affect phenotype by altering expression of protein coding genes. The extent of its contribution to phenotypic diversity over evolutionary time is, however, unclear, because of limited stability of epialleles that are not linked to DNA mutations. To dissect the relative contribution of DNA methylation to transposon surveillance and host gene regulation, we leveraged information from three species in the Brassicaceae that vary in genome architecture, *Capsella rubella*, *Arabidopsis lyrata*, and *Arabidopsis thaliana*. We found that the lineage-specific expansion and contraction of transposon and repeat sequences is the main driver of interspecific differences in DNA methylation. The most heavily methylated portions of the genome are thus not conserved at the sequence level. Outside of repeat-associated methylation, there is a surprising degree of conservation in methylation at single nucleotides located in gene bodies. Finally, dynamic DNA methylation is affected more by tissue type than by environmental differences in all species, but these responses are not conserved. The majority of DNA methylation variation between species resides in hypervariable genomic regions, and thus, in the context of macroevolution, is of limited phenotypic consequence.

Contributions

Conceived and designed the experiments: DKS DK CB DW. Performed the experiments: DKS DK. Analyzed the data: DKS DK JH CB. Contributed to the writing of the manuscript: DKS DK DW.

License

“PLOS applies the Creative Commons Attribution (CC BY) license to works we publish. This license was developed to facilitate open access – namely, free immediate access to, and unrestricted reuse of, original works of all types. Under this license, authors agree to make articles legally available for reuse, without permission or fees, for virtually any purpose. Anyone may copy, distribute or reuse these articles, as long as the author and original source are properly cited.”

<http://journals.plos.org/plosone/s/content-license>



Evolution of DNA Methylation Patterns in the Brassicaceae is Driven by Differences in Genome Organization

Danelle K. Seymour¹, Daniel Koenig¹, Jörg Hagmann, Claude Becker, Detlef Weigel*

Department of Molecular Biology, Max Planck Institute for Developmental Biology, Tübingen, Germany

Abstract

DNA methylation is an ancient molecular modification found in most eukaryotes. In plants, DNA methylation is not only critical for transcriptionally silencing transposons, but can also affect phenotype by altering expression of protein coding genes. The extent of its contribution to phenotypic diversity over evolutionary time is, however, unclear, because of limited stability of epialleles that are not linked to DNA mutations. To dissect the relative contribution of DNA methylation to transposon surveillance and host gene regulation, we leveraged information from three species in the Brassicaceae that vary in genome architecture, *Capsella rubella*, *Arabidopsis lyrata*, and *Arabidopsis thaliana*. We found that the lineage-specific expansion and contraction of transposon and repeat sequences is the main driver of interspecific differences in DNA methylation. The most heavily methylated portions of the genome are thus not conserved at the sequence level. Outside of repeat-associated methylation, there is a surprising degree of conservation in methylation at single nucleotides located in gene bodies. Finally, dynamic DNA methylation is affected more by tissue type than by environmental differences in all species, but these responses are not conserved. The majority of DNA methylation variation between species resides in hypervariable genomic regions, and thus, in the context of macroevolution, is of limited phenotypic consequence.

Citation: Seymour DK, Koenig D, Hagmann J, Becker C, Weigel D (2014) Evolution of DNA Methylation Patterns in the Brassicaceae is Driven by Differences in Genome Organization. PLoS Genet 10(11): e1004785. doi:10.1371/journal.pgen.1004785

Editor: Brandon S. Gaut, University of California Irvine, United States of America

Received: July 18, 2014; **Accepted:** September 26, 2014; **Published:** November 13, 2014

Copyright: © 2014 Seymour et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability: The authors confirm that all data underlying the findings are fully available without restriction. Raw reads are deposited at the European Nucleotide Archive (<http://www.ebi.ac.uk/ena/>) under accession number PRJEB6701.

Funding: Supported by the Deutsche Forschungsgemeinschaft (DFG) (SPP1529 Adaptomics), a Human Frontier Science Program (HFSP) Long-Term Fellowship (DK), and the Max Planck Society. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

* Email: weigel@weigelworld.org

† These authors contributed equally to this work.

Introduction

Cytosine methylation is a heritable epigenetic modification found in the genomes of organisms spanning the eukaryotic phylogeny [1,2,3,4]. It occurs in three nucleotide contexts, CG, CHG, or CHH (where H is any nucleotide except G) [5], and is enriched in the repeat rich heterochromatic regions of genomes, in nucleosome linkers, and at CG sites in the exon sequences of genes (gene body methylation) [4,6,7,8,9,10,11]. Repeat-localized DNA methylation plays a role in transposon silencing [12,13], but the direct relationship between transcription of protein coding genes and DNA methylation remains unclear. In contrast to repeat methylation, gene body methylation is associated with moderately transcribed sequences [6,7,14,15,16], and has been proposed to stabilize gene expression levels by excluding H2A.Z [17]. Nevertheless, DNA methylation can vary between tissues and environments [18,19,20], and in a handful of cases changes in methylation state contribute to heritable phenotypic variation, although the majority have been linked to structural differences near the affected genes [21,22,23,24,25,26,27]. These observations suggest that DNA methylation may regulate developmental processes and that it could potentiate phenotypic variation during evolution.

Unlike mutational processes acting on DNA sequences, our understanding of the factors contributing to meiotically stable variation in DNA methylation is in its infancy [28]. The different molecular mechanisms governing DNA methylation constitute one factor impacting stability and subsequent inheritance at symmetric and asymmetric sites. In the plant *Arabidopsis thaliana*, initiation and maintenance of methylation at CG and CHG sites is divided primarily between DNA METHYLTRANSFERASE 1 (MET1) and CHROMOMETHYLASE3 (CMT3) [29,30,31]. During DNA replication these two enzymes copy symmetrically methylated cytosines onto the newly synthesized DNA strand using the parental strand as a template [32,33]. Unlike symmetric cytosine methylation, CHH methylation cannot be replicated from the template strand [34]. Instead, methylation at newly synthesized CHH sites is established after cell division by the RdDM RNA-directed DNA methylation pathway through the concerted action of small RNAs (sRNAs) produced from the methylated locus and the de novo DNA methyltransferases DRM1/DRM2 (DOMAINS REARRANGED METHYLTRANSFERASE1/2) [34,35,36,37]. In addition, RdDM-independent asymmetric DNA methylation relies on DDM1 (DECREASE IN DNA METHYLATION1) and CMT2 [38].

Author Summary

DNA methylation is an epigenetic mark that has received a great deal of attention in plants because it can be stably transmitted across generations. However, the rate of DNA methylation change, or epimutation, is greater than that of DNA mutation. In addition, different from DNA sequence, DNA methylation can vary within an individual in response to developmental or environmental cues. Whether altered characters can be passed on to the next generation via directed modifications in DNA methylation is a question of great interest. We have compared how DNA methylation changes between species, tissues, and environments using three closely related crucifers as examples. We found that DNA methylation is different between roots and shoots and changes with temperatures, but that such changes are not conserved across species. Moreover, most of the methylated sites are not conserved between species. This suggests that DNA methylation may respond to immediate fluctuations in the environment, but this response is not retained over long evolutionary periods. Thus, in contrast to transcriptional responses, conserved epigenetic responses at the level of DNA methylation are not widespread. Instead, the patterns of DNA methylation are largely determined by the evolution of genome structure, and responsive loci are likely short-lived accidents of this process.

The extent to which DNA methylation varies at individual sites across generations, or the epimutation rate, has only recently been characterized in isogenic plant lines [39,40]. Repeat-associated methylation was remarkably stable over 30 generations, but some variability arose outside of repeats in euchromatic sequence [39,40]. Changes in DNA methylation accumulated non-linearly, indicating that a subset of methylated sites is particularly prone to spontaneous changes in methylation and, as a result, the absolute DNA methylation differences quickly reach saturation [39,40]. Variation of methylation across generations has been linked to the transgenerational cycling of transposon and repeats between methylated and unmethylated states in the germline [41].

Armed with the knowledge of within-species epimutation rate, the degree of epigenome stability over short evolutionary periods, within a single species, for example, can be addressed [18]. Using *A. thaliana*, intraspecific variation in methylation was surveyed in 140 geographically diverse accessions [18]. Most single site and RdDM-derived regional epimutations were rare, occurring in only a few of the 140 accessions [18]. The lack of intermediate frequency epimutations in these categories is consistent with the view that the vast majority of new methylation variants within a species may only exist for brief periods during evolution. Not too surprisingly, a significant subset of both rare and intermediate frequency RdDM-derived regional epimutations were associated with previously unknown structural variants [18]. Expansion and contraction of repeat-associated sequences leads to intraspecific structural variation; therefore, as a result of RdDM silencing, such structural variants should be linked to methylation variation.

Over longer evolutionary periods, broad similarities in DNA methylation are observed across a variety of genomic features. Large-scale patterns of methylation are shared across flowering plants, including extensive methylation of heterochromatic transposon and repeat-associated sequences [6,7,8,9,10,11] likely due to conservation of the RdDM machinery in plants. Over shorter divergence times, similar levels of gene body methylation have been observed at orthologous genes within the grasses [11,42]. Similarly, in vertebrates, where most of the CG sites in the genome

are methylated, absence of methylation at so-called CpG islands is usually found in all species examined [43]. Regardless of organism, the degree of DNA methylation conservation depends on both the evolutionary time scale under consideration and on the genomic feature of interest.

Here we compare at single base resolution DNA methylation in three closely related Brassicaceae - *Capsella rubella*, *Arabidopsis lyrata*, and *Arabidopsis thaliana*. These three species, which diverged about 10 to 20 million years ago [44], vary in genome size and architecture [45,46,47]. Both *C. rubella* and *A. lyrata* have a Brassicaceae typical set of eight chromosomes, while *A. thaliana* has only five chromosomes [48,49]. Both the *A. lyrata* and *C. rubella* genomes are about 50% larger than that of *A. thaliana*, but for very different reasons. Expansion of centromeric, heterochromatic regions has enlarged the *C. rubella* genome, but predominantly euchromatic regions have expanded in *A. lyrata*, driven by insertions of transposable elements (TEs) adjacent to genic sequences [46,47]. Reflecting these differences in genome architecture, the reference genome assemblies represent about 85% of the entire genome in *A. lyrata*, about 75% in *A. thaliana*, and about 60% in *C. rubella* (Table S1) [46,47,50,51,52,53,54]. We show that the difference in genome structure is a major factor influencing the evolution of DNA methylation in these species. Furthermore, while overall DNA methylation is similar between species at many sites, dynamic DNA methylation responses between environments and tissues are rarely conserved. Using a comparative framework we were able to disentangle the contribution of genomic, environmental, and developmental factors to DNA methylation variation between species.

Results

Genome-wide distribution of DNA methylation

Using a factorial design, we subjected seedlings of the inbred reference strains, *A. thaliana* Col-0, *A. lyrata* MN47, and *C. rubella* MTE, to either a control or 23-hour cold treatment and separately harvested root and shoot tissues. This design provides the opportunity to determine conservation of DNA methylation as well as dynamic changes between and within species. In addition to extracting DNA for bisulfite-sequencing in duplicate, we also extracted RNA in triplicate for RNA-seq.

Bisulfite-treated samples were sequenced to an average of 20× strand-specific coverage (Table S2). With this coverage, over 97.5% of the cytosines in the non-repetitive portion of the reference genome of each species could be interrogated (99.5% for *C. rubella*, 97.5% for *A. lyrata*, and 98.7% for *A. thaliana*). With a minimum coverage of three, we confidently estimated methylation rates at two thirds to three quarter of cytosines (62% for *C. rubella*, 65% for *A. lyrata*, and 75% for *A. thaliana*). Sites with significant methylation levels were identified using a binomial test [39]. False positive rates, determined from incomplete conversion of exogenous unmethylated phage lambda DNA, were very low (Table S3).

Global patterns of DNA methylation in *A. lyrata* and *C. rubella* are similar to those reported before for *A. thaliana*, with highest levels in regions near the centromeres, which are populated by TEs and repeats, but contain few genes [6,14,15] (Fig. 1). There is little correlation between DNA methylation density and gene expression at the 500 kb scale (Fig. 1). Centromeric regions are plagued with TEs, and as expected, methylation is found preferentially at sites annotated as residing in TEs (Fig. 2A). Methylation at CHG and CHH sites, which account for over half of methylated sites in all three species, occurs almost exclusively in TEs (Fig. 2A).

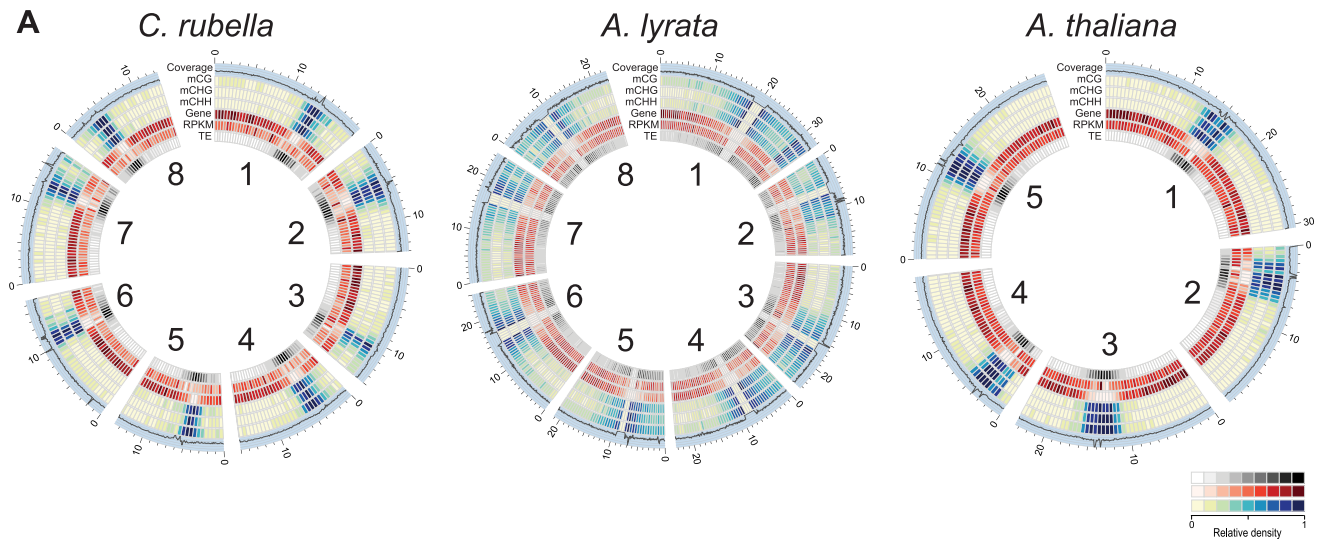


Figure 1. Genomic distribution of DNA methylation. A) Circos plots [74] of *C. rubella*, *A. lyrata*, and *A. thaliana*. Chromosome number is indicated on the inner circle. Data is plotted for 500 kb windows, except for sequencing coverage (100 kb). Gene expression (RPKM) was calculated using the sum of the expression counts from all samples within a species. doi:10.1371/journal.pgen.1004785.g001

Methylation patterns in the three species reflect their genome architecture. While we mapped a similar number of methylated cytosines in *A. thaliana* and *C. rubella*, consistent with the almost equal size of euchromatic sequences in both species, we identified almost three times as many methylated cytosines in *A. lyrata*, even though its reference genome assembly is only 50 to 75% longer than that of the other two species. The larger number of methylated cytosines in *A. lyrata* has led to an elevation in the methylation rate at a number of genomic features (Fig. 2B). This increase has only occurred at CHG and CHH sites, hallmarks of RdDM at TEs, and is especially evident in introns, correlating with the invasion of introns by TEs in this species (Fig. 2B, C). Almost one third of intronic bases in *A. lyrata* overlap with a TE or repeat, compared to fewer than 10% in the other two species (Fig. 2C), with the expansion found for all TE classes (Fig. 2D). Intron-inserted TEs are frequently found in non-expressed genes (Fig. S1) and are associated with increased methylation in flanking intronic and exonic sequences (Fig. S2), potentially due to pseudogenization or incomplete annotation of repeats. However, when a TE is inserted into the intron of an expressed gene, elevation of CHG and CHH methylation of exon sequences is not evident (Fig. S2, S3). Despite TE expansion in *A. lyrata*, the level of *A. lyrata* gene body methylation is comparable to that of *C. rubella*, which has few TEs in its introns (Fig. 2E). However, species-specific differences in methylation patterns are evident in flanking UTR and intergenic sequence (Fig. 2E). In these regions *A. lyrata* is the most highly methylated in all contexts (Fig. 2E). Depending on context, *C. rubella* displays methylation levels either similar to *A. thaliana* or intermediate between the two other species (Fig. 2E).

Arabidopsis thaliana lost three centromeres relative to *A. lyrata* and *C. rubella*, and this loss has been estimated to account for about 10% of the genome size reduction in *A. thaliana* [46]. Using orthologous genes, it is possible to reconstruct the gene, repeat, and methylation density using the ancestral chromosome positions (Fig. 3). As expected, repeat density and cytosine methylation next to these degraded centromeres is reduced in *A. thaliana*, while gene density is higher (Fig. 3). Particularly notable is the decrease in CG gene body methylation (Fig. 3). Although gene body methylation is positively correlated with gene expression in several

species [6,7,14,15,16], gene expression is not noticeably different in these regions between the three species (Fig. 3). Thus, the elimination of centromeres has had a measurable impact on repeat and methylation distribution in *A. thaliana*, but did not strongly affect the expression of ancestrally pericentromeric genes.

Methylated regions are not conserved across species

Methylation of plant genomes is driven to a large extent by TEs, which are silenced via either the sRNA-mediated RdDM pathway [36] or the RdDM-independent pathway which relies on DDM1 [38]. Using a Hidden Markov Model algorithm, we identified methylated regions (MR) in each genome, which have a median length of 300 to 530 bp and cover between 26 and 73 Mb (Table S4). MRs are preferentially found in heterochromatic sequence next to centromeres, as they are enriched for TEs (Fig. S4, Fig. 4A). Since TEs are rapidly turned over, we expected MRs to be only poorly conserved. To test this assumption, we identified nearly 60 Mb of sequences with a 1:1:1 relationship in whole-genome alignments (Table S5) [47]. Less than 1% of the MR space is contained in the alignable portion of the genomes (Fig. 4B). In the rare cases where an MR spans alignable sequences, such sequences are almost always methylated in only one of the three species (Fig. 4C). We conclude that DNA methylation targets primarily the variable portion of the genome, which is subject to species-specific expansion and contraction of TEs.

To determine whether specific orthologs tend to be associated with methylation in all species, even in the absence of MR sequence conservation, we analyzed orthologs that contained a MR overlapping or within 1 kb of their coding region. Again, we found that the presence of MRs is rarely conserved (Fig. 4D, Table S6), although MR sharing is seen more often than expected by chance (Fig. 4D, Tables S7, S8). This could, however, be simply due to genes near centromeres being more often associated with MRs because they are in an MR-rich genome environment.

Conservation of CG gene body methylation

In contrast to RdDM of TEs and other repeats, the function of CG gene body methylation is still enigmatic, although it correlates

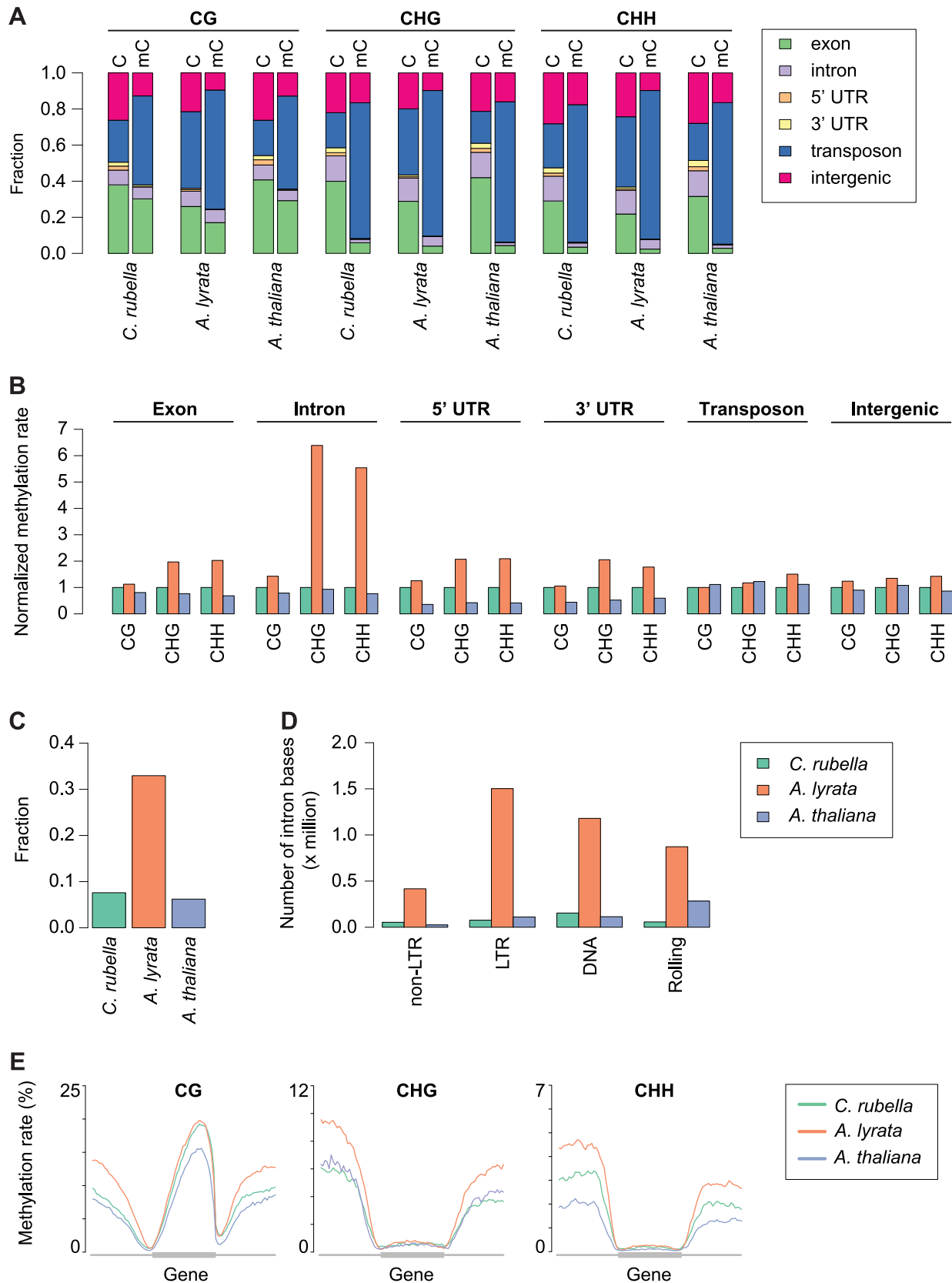


Figure 2. Impact of repeat expansion on DNA methylation at genomic features. A) Feature annotation of all cytosines and methylated cytosines. Annotations are shown for all three contexts. B) Genome average of methylation rates for each genomic feature. Methylation rates are normalized to the outgroup species *C. rubella*. C) Fraction of intron bases annotated as transposable element or other repeat sequence. D) Total number of intron bases (millions) that are annotated as a particular transposable element class. E) Methylation rate distribution across gene bodies of

orthologous genes and flanking sequences (1.5 kb up- and downstream). Orthologs that lacked methylation in both their gene body and flanking sequences were excluded. Distributions are plotted by context. doi:10.1371/journal.pgen.1004785.g002

positively with gene expression and negatively with mean normalized expression variance, or the coefficient of variation, across tissues and treatments (Fig. S5) [6,7,14,15,16,17]. CG gene body methylation is found in the majority of genes (Table S9), and its rate is highly correlated between orthologs, while CG methylation up- and downstream of genes is much less correlated (Fig. 5).

CHG and CHH methylation in gene bodies is often indicative of transcriptionally inactive pseudogenes, paralogs, or transposons wrongly annotated as protein coding genes [14,15,55]. Between 10 and 20% of genes exhibit CHG or CHH methylation, most of which were not expressed in our samples (Table S9). Genes with CHG or CHH methylation are underrepresented in the orthologous gene set, where their fraction drops to less than half of their fraction among all genes, supporting the assertion that

CHG and CHH methylation point to a tendency toward pseudogenization (Table S9). Moreover, CHG and CHH methylation are generally not conserved, suggesting that these marks arise in a lineage-specific fashion.

Site-specific gains and losses of methylation in euchromatic sequence

We used the cross-species alignments to identify 15.1 million conserved CG, CHG and CHH sites, which are located particularly in exons (Fig. 6A, Table S5). Although only a small portion, 2%, had significant methylation, most were shared between at least two species, with *A. thaliana* having the fewest methylated sites, reflecting the general decrease in global DNA methylation in this species (Fig. 6B–D, Table S10). Sites

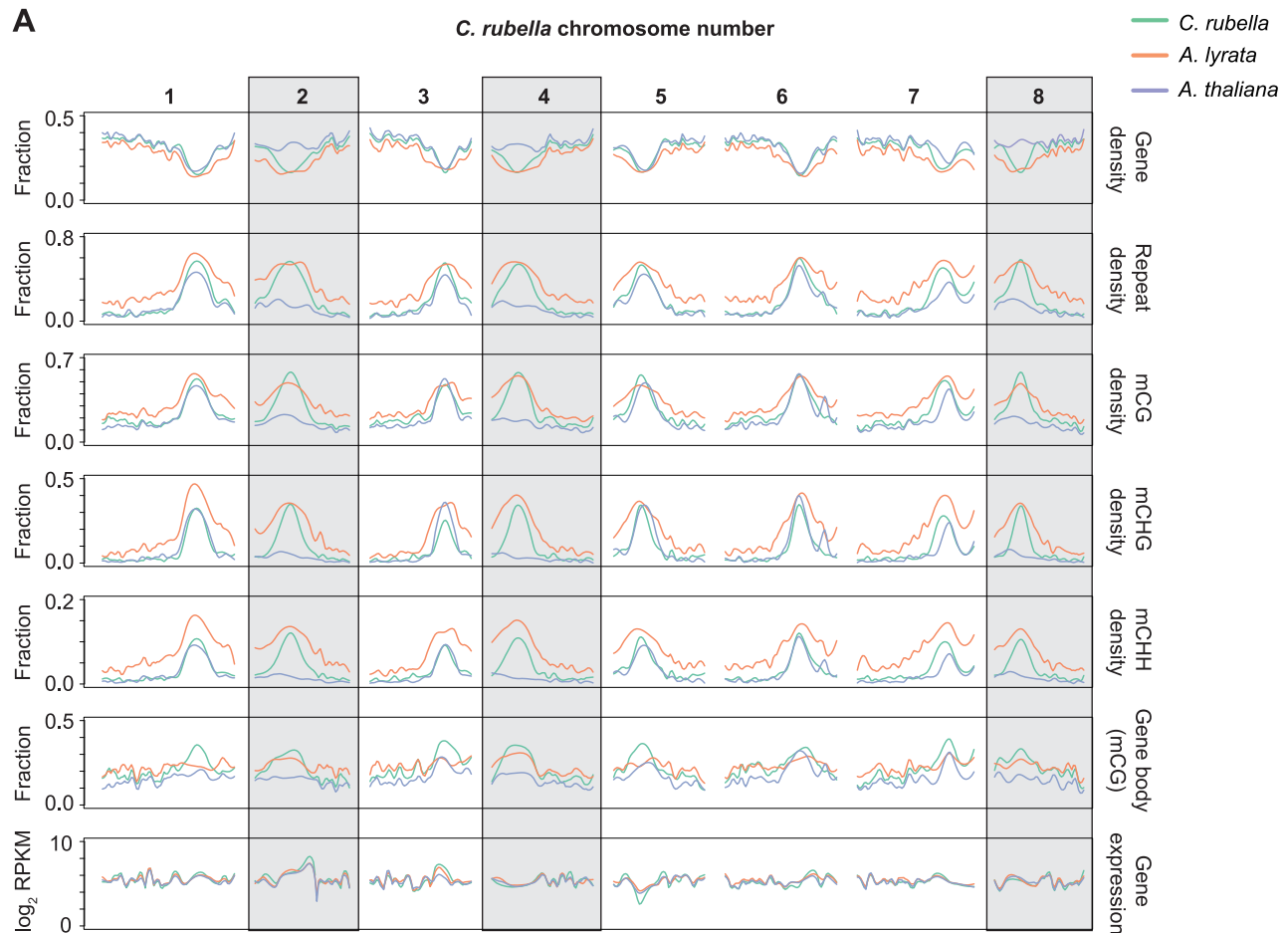


Figure 3. Centromere loss impacts DNA methylation in *A. thaliana*. A) Orthologous genes, anchored on the *C. rubella* genome, were used to calculate several statistics to investigate the impact of centromere loss on DNA methylation in *A. thaliana*. *Capsella rubella* centromeres 2, 4, and 8 (grey boxes) were lost during chromosomal fusion events that occurred on the branch leading to *A. thaliana*. Gene density, repeat density, and methylation densities were calculated for a 20 Kb window centered on the midpoint of each orthologous gene (10 kb up- and 10 kb downstream). Gene density and repeat density were calculated as fractions of each 20 kb window annotated as either a gene (ATG to STOP) or a repeat. Methylation densities were calculated as fractions of cytosines methylated in each context. Gene body methylation and gene expression (RPKM) were calculated for each ortholog. Gene body methylation was calculated as the fraction of methylated CG sites in a gene (ATG to STOP). Gene expression data from all samples within a species were used to calculate the RPKM values. For each statistic, local linear regression was performed to smooth the data in 250 kb bins. Smoothing parameter was relative to chromosome length. doi:10.1371/journal.pgen.1004785.g003

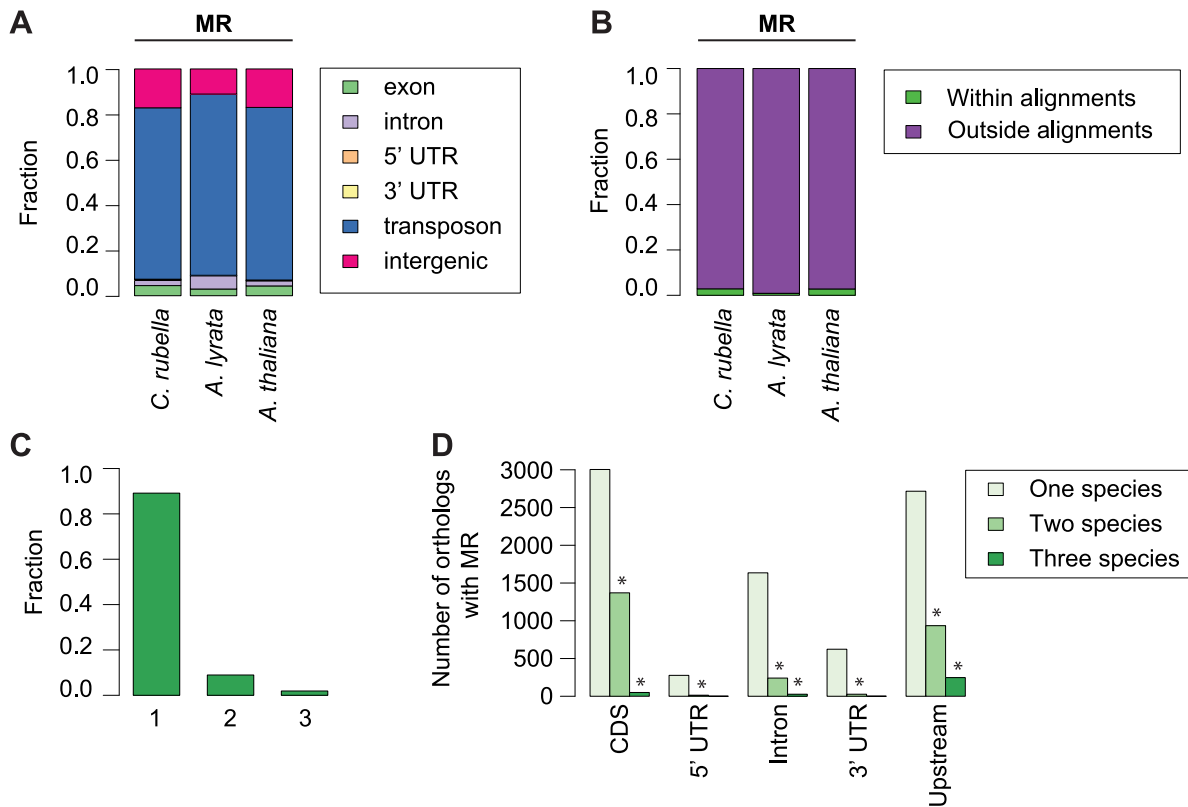


Figure 4. Conservation of methylated regions (MR). A) Annotation of all bases in MRs. B) Fraction of bases in MRs that occur either within or outside of the three-way whole genome alignments. C) Fraction of MR bases found within three-way whole genome alignments that occur in one, two, or three species. D) Conservation of MRs in the absence of sequence alignments. The total number of orthologous genes overlapping an MR in one, two, or three species is given, with location of MR overlap separated by genomic feature. Upstream region was defined as 1 kb before the start codon. Asterisk indicates two or three-way sharing of MRs that exceeds permutation values. doi:10.1371/journal.pgen.1004785.g004

methylated in multiple species are further enriched in exons, with very few of these conserved sites being CHG or CHH sites (Fig. 6B,C, Fig. S6).

Sites that differ in methylation between species can be used to study gain and loss of methylation. We consider sites that are methylated only in a single species as lineage-specific gains, and absence of methylation in only one species as lineage-specific losses. We found that the number of gains and losses reflect the differences in genome architecture between the three species (Fig. 6 B,D). The many methylation losses in *A. thaliana* appear to be the result of genome shrinkage, and this species has also the fewest gains. In contrast, *A. lyrata* has the most gains, likely reflecting recent TE expansion (Fig. 6 B,D). The density of variable sites across the genome (in 10 kb windows) illustrates that gains and losses are not randomly distributed (Fig. 6D). Species-specific gains, which occur in all three sequence contexts, are concentrated in a subset of windows that are strongly enriched for TEs (Fig. 6D,E), but are also frequently found in exons (Fig. S6). That methylation gains are particularly likely in first and last exons suggests that methylation spreading from nearby TEs makes an important contribution to newly methylated sites, regardless of TE class (Fig. 6F, S7) [56,57,58].

Lineage-specific losses are more evenly distributed, without any signature of TE association. In addition, sites that are conserved in not only two, but all three species occur across a similar spectrum of genomic features (Fig. S6). Together these results indicate that unlike gains, losses occur in a random fashion, with the proviso

that there is an overall global loss of methylation in *A. thaliana* (Fig. 6D). Though centromere elimination contributes to the different methylation pattern in *A. thaliana*, this explains only a minority of these losses (Fig. S8). It appears more likely that they are caused by the global reduction in TE content. We also attempted to understand what factors might contribute to conservation of DNA methylation over time. Sites found in more than one species are enriched in exons of conserved length and are more frequent in the center of exons (Fig. S9, S10).

Methylation variation within individuals

Because several studies have shown that DNA methylation can change between tissues and in response to external stimuli [19,20], we wanted to address whether these responses are conserved. Principal component analysis on the four types of samples, control shoots, cold-treated shoots, control roots and cold-treated roots, for all three species according to global RNA-seq measurements revealed that tissue is the most important factor, with over 7,000 genes being differentially expressed between roots and shoots (Fig. 7A, S11). Tissue-specific differences in gene expression are the largest source of expression variance in this data set (Fig. 7A). In contrast, species is the most important factor for differences in DNA methylation and explains 80% of the variance in our data (Fig. 7B, Fig. S12). Moreover, PC2 places *A. lyrata* closest to *C. rubella* instead of its congener *A. thaliana*, reflecting the methylation losses in *A. thaliana* (Fig. 7B).

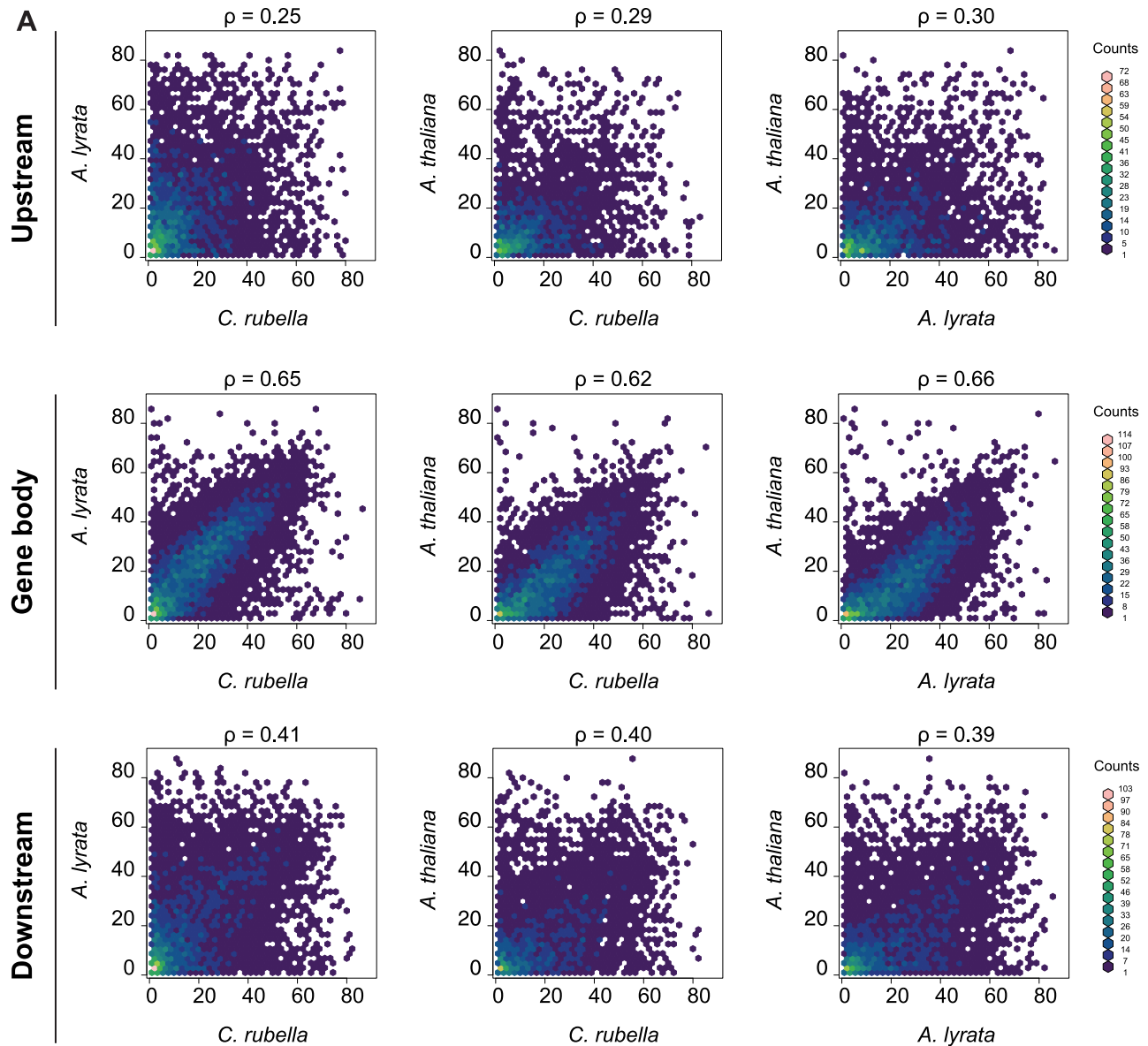


Figure 5. Methylation rates at orthologs. A) Pairwise comparison of the average methylation rates at orthologs. Average methylation rate was calculated as the average of all CG sites in the feature, including non-methylated CG sites. Pairwise comparisons are shown for upstream regions (1.5 kb), gene bodies, and downstream regions (1.5 kb). Spearman rank correlation coefficient (ρ) is included for each comparison. doi:10.1371/journal.pgen.1004785.g005

To evaluate the degree to which within-species DNA methylation changes are conserved, we first estimated significant differential methylation at site and region levels. Four biologically appropriate comparisons were performed for each species to minimize multiple testing problems. Two tests identified differentially methylated positions (DMPs) between roots and shoots, and two tests identified DMPs between cold and control conditions regardless of tissue type. In each species, ten times as many DMPs were found between tissues than between treatments (Figure 8A, Table S11). Similar to DMPs, 20 to 50 times as many differentially methylated regions (DMRs) were detected between tissues than between treatments (Fig. 8B, Table S12).

Importantly, DMPs and DMRs do not necessarily coincide (Fig. S4, S13). DMPs in all contexts are rarely found within DMRs, indicating that significant regional changes in methylation are not just the extension of single base differences (Fig. 8C). CHG and

CHH DMPs reside mainly within MRs (Fig. 8C); since these are almost exclusively found in the non-alignable portions of the genome, including TEs (Fig. 4A, Fig. 8D), the positions of DMPs and DMRs are typically not conserved between species (Fig. 8E). In the rare case that DMPs or DMRs can be found in the portion of a species' genome that can be aligned with the genomes of the other two species (Fig. 8E), they are only variable in a single species (Fig. 8F). Methylation variation at both the site and region level is therefore not conserved across species.

In the absence of sequence conservation at DMRs, we looked for conservation of their presence at orthologous genes. When only considering orthologs, fewer than 700 genes coincide with a DMR (405 in *C. rubella*, 652 in *A. lyrata*, and 221 in *A. thaliana*) (Table S13). Orthologs only rarely shared the presence of an overlapping or adjacent DMR, similar to what we see for MRs. Despite the rarity of such cases, they occur more often than expected by

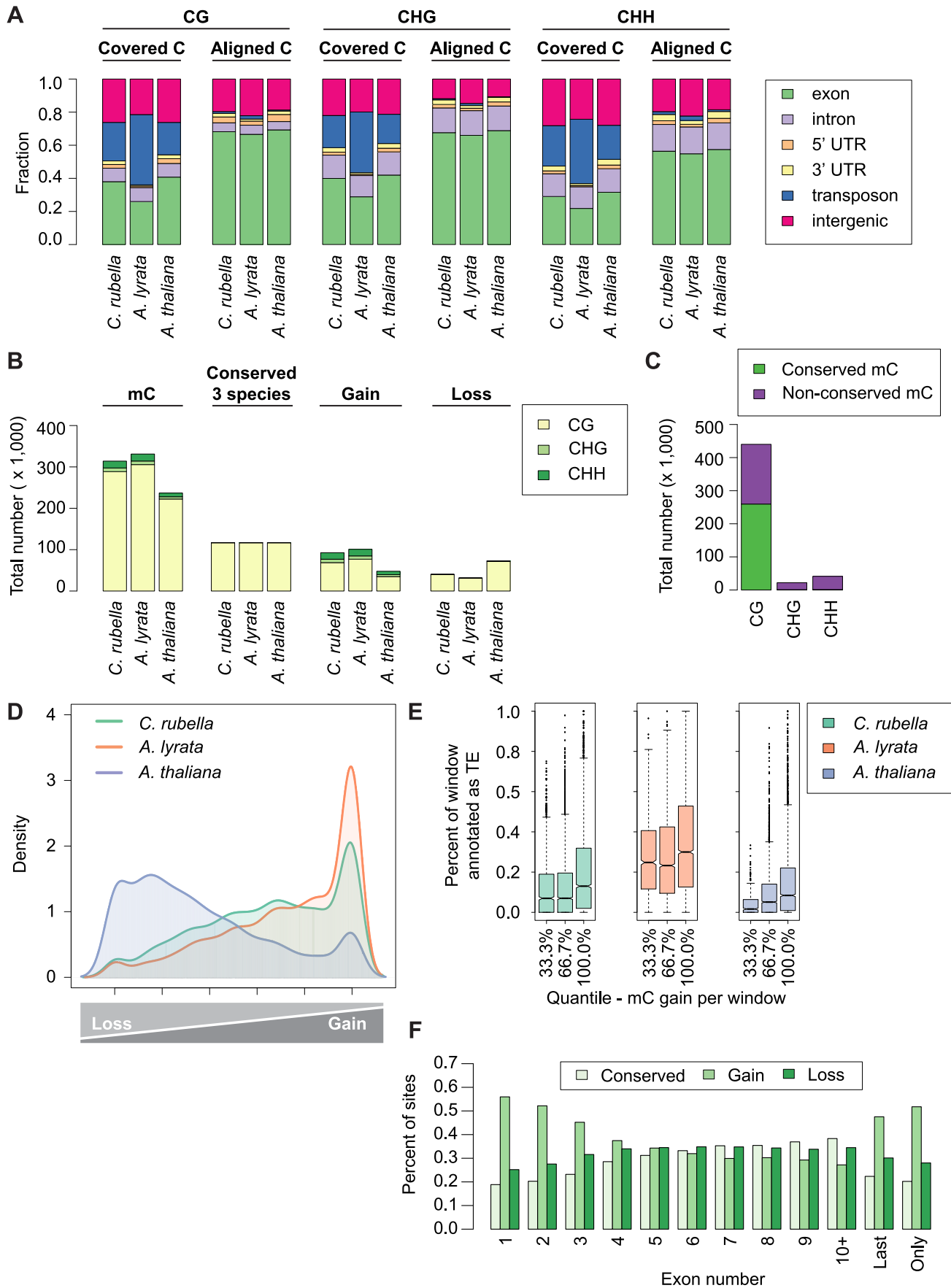


Figure 6. Site-level comparison of methylation. A) Annotation of all cytosines within a species (covered C) compared to the annotation of cytosines found in the three-way whole genome alignments (aligned C). B) Total number of mC by context for aligned site classes. Site classes are as follows: mC - methylated sites within a species. Conserved (3 species) - sites that are methylated in all three species. Gain - sites that are methylated in a single species. Loss - sites that have lost methylation in a single species. C) Total number of conserved mC and non-conserved mC by context. D) Density plot describing the distribution of variable sites in the genome (10 kb windows). For each window the following statistic was calculated: species-specific methylation gains/sum of species-specific methylation gains and losses. E) Windows with a high density of gains have more transposons and repetitive sequences. Density of transposons plotted against density of methylation gains (10 kb window). F) Methylation gains are enriched at the beginning and end of genes. Fraction of mC in each site class is plotted by exon position in a gene.
doi:10.1371/journal.pgen.1004785.g006

chance for a subset of genomic features and species comparisons (Fig. S14, Table S14, Table S15). Lack of sequence conservation together with minimal overlap of DMR presence at orthologs supports the transitory nature of methylation variation during genome evolution.

We also asked whether differential methylation in or near coding sequences is correlated with changes in gene expression. DMP and DMR overlap with genes was analyzed separately for those that overlapped with exons, introns, 5' UTRs, 3' UTRs and 1 kb upstream regions (Table S13, S16). DMPs occur in many genes in all three species, and most of them are expressed in our samples (9,631 in *C. rubella*, 12,216 in *A. lyrata*, and 6,345 in *A. thaliana*), but there is no evidence for correlation between DMPs and gene expression. This holds true for tissue as well as treatment DMPs (average Spearman rank correlation coefficient tissue = -0.04, treatment = 0.02, Table S17). Only a small number of DMRs overlap with expressed genes (529 in *C. rubella*, 801 in *A. lyrata*, and 284 in *A. thaliana*). Again, there is no correlation with gene expression (average Spearman rank correlation coefficient for CG DMRs = -0.16, CHG DMRs = -0.06, CHH DMRs = 0.00, Table S18).

Although DMPs and DMRs are not conserved across species, there is consistently more variability between root and shoot samples at a number of genomic features. Importantly, the methylation profile across transposons is quite different between

tissues. Transposons are consistently more highly methylated in all sequence contexts in shoots (Fig. 9A). A similar trend is apparent for CHG and CHH sites in intergenic regions in *A. lyrata*, reflecting that TEs are closer to genes in this species (Fig. 9B) [46].

Discussion

DNA methylation is an ancient epigenetic modification that appears in the genomes of organisms throughout the eukaryotic phylogeny [1,2,3]. This mark is associated with a number of cellular processes including transposon silencing and host gene regulation, but the cause-and-effect relationship between gene expression and DNA methylation remains unclear [6,7,12,13,14,15,16]. From an evolutionary standpoint, it is useful to consider methylated cytosines from two differing perspectives, either as a non-canonical nucleotide or as a molecular phenotype akin to transcription, and each perspective has important implications for the interpretation of its evolutionary dynamics.

Dynamics of DNA methylation as a molecular phenotype

As a molecular phenotype, many characteristics of DNA methylation are conserved between the species we examined. DNA methylation is generally associated with the repeat-dense sequences found in the centromeres, with CG methylation being in addition present at high levels in exonic sequences

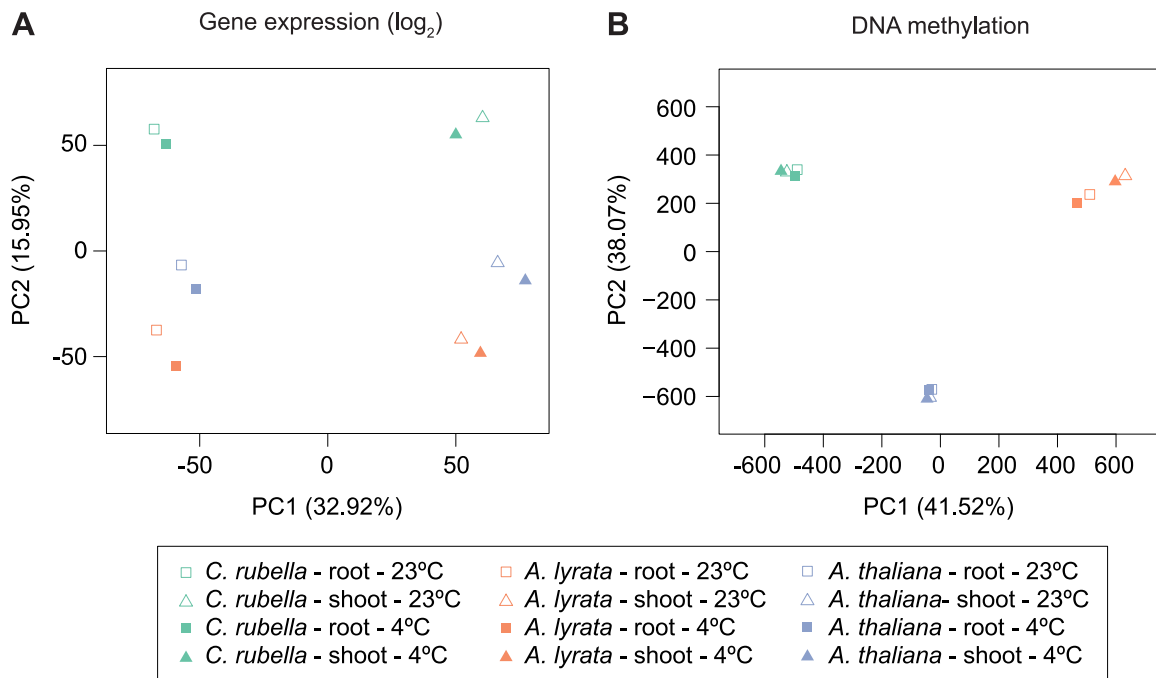


Figure 7. Species gene expression and mC relationships. A) Principal component analysis on fitted gene expression values (\log_2) and B) mC rates at aligned methylated positions. All contexts are considered (see Fig. 6B,C and Table S10 for further description of mC sites).
doi:10.1371/journal.pgen.1004785.g007

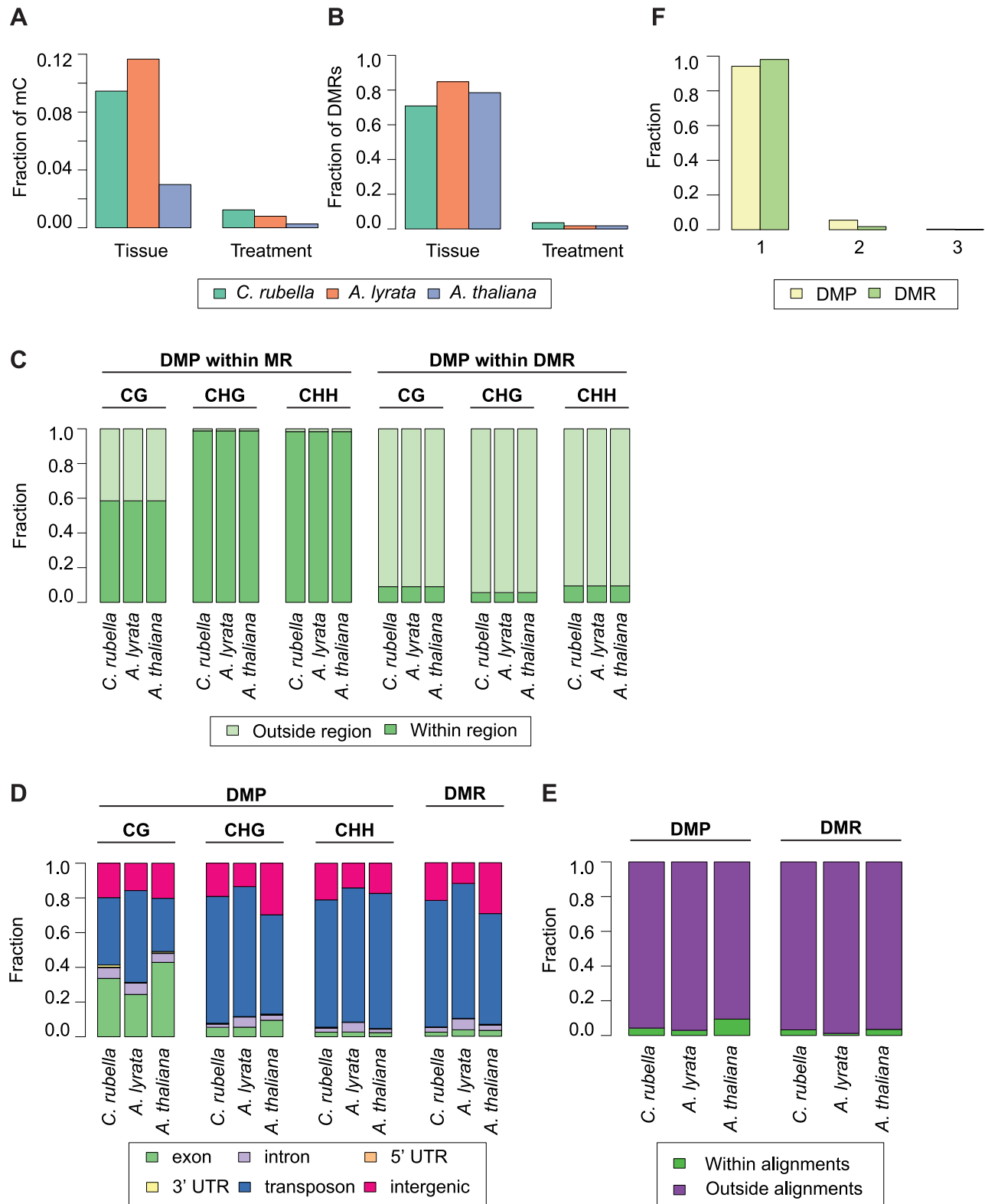


Figure 8. Intraspecific variation in DNA methylation. A) Fraction of mC that are variable between either tissue (root and shoot) or treatment (23°C and 4°C) comparisons. B) Fraction of DMRs that are variable between either tissue (root and shoot) or treatment (23°C and 4°C) comparisons. C) Fraction of DMPs in each context that reside either within a MR or DMR. D) Feature annotation of DMPs by context and DMR bases. E) Fraction of DMPs and DMR bases found within three-way whole genome alignments. F) Fraction of DMPs and DMR bases found within three-way whole genome alignments that occur in one, two, or three species.

doi:10.1371/journal.pgen.1004785.g008

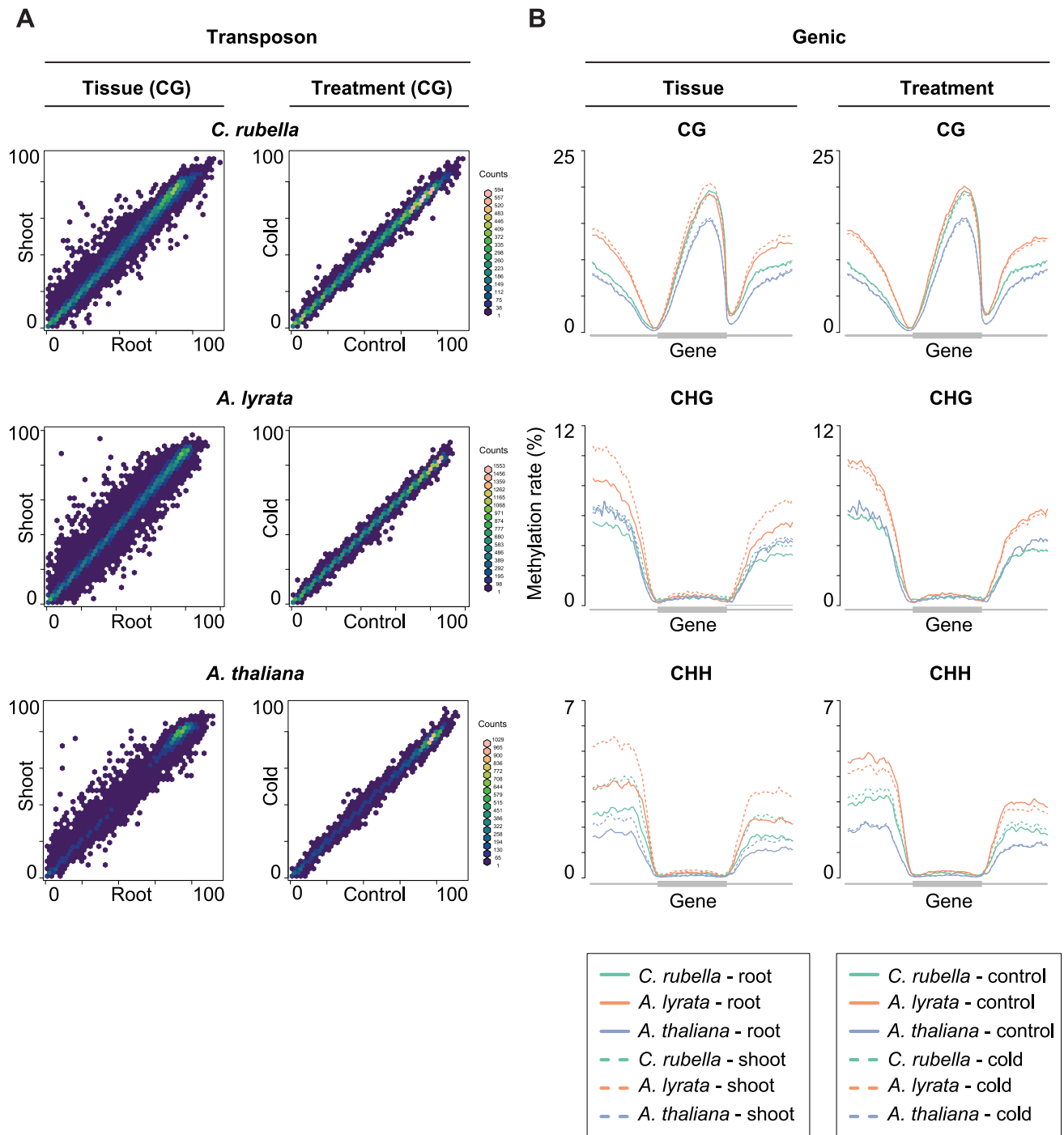


Figure 9. Intraspecific variation of transposon and gene body methylation. A) Comparison of the average methylation rates at annotated transposons and repeats between tissues (root and shoot) and treatments (23°C and 4°C). Average methylation rate is calculated as the average of methylation rates at all cytosines in the feature, including non-methylated cytosines. B) Methylation rate distribution across gene bodies of orthologous genes and flanking sequences (1.5 kb up- and 1.5 kb downstream). Orthologs that lacked methylation in both their gene body and flanking sequences are excluded. Distributions are plotted by context. doi:10.1371/journal.pgen.1004785.g009

[6,7,8,9,10,11,14,15,16]. Furthermore, gene body methylation levels are conserved in orthologous genes indicating that DNA methylation rate may be subject to purifying selection, a finding consistent with previous wider evolutionary comparisons [42]. The close relationship of the species used in our experiments allows us to make inferences at base pair resolution. Given the substantial

rate of epimutation in non-repetitive sequences [39,40], we were surprised to discover that a large fraction of sites is methylated in more than one species. These sites were predominantly found in gene bodies, providing additional evidence for selective constraint. While gene body methylation is poorly understood, there is some evidence that it is correlated with nucleosome positioning in exons

[14,59]. If nucleosome position is conserved, it could potentially explain long-term conservation of DNA methylation at some sites.

An additional proposed feature of DNA methylation as a molecular phenotype is the ability to respond to external stimuli or internal developmental cues. In theory, such variation could control changes in gene expression. We found evidence for DNA methylation variation in all three species across both tissue type and environment. The changes in DNA methylation were in all three species much greater between tissues, and consistently resulted in lower methylation levels in the root [19]. Differences between the root and shoot tissues also explain a majority of the expression variation in the transcriptional data, but these changes are not directional. We found no evidence that changes in DNA methylation across tissues is associated with changes in gene expression. In fact, a large proportion of methylation changes were found in repetitive sequences. This pattern may result from the increased stringency of transposon silencing in the shoot, which includes the plant germline [60].

While transcriptional responses are highly conserved across all three species, we found no evidence for conservation of DNA methylation response at the sequence level. MRs and DMRs are predominantly found in the rapidly evolving repeat-rich regions of the genome and rarely reside in or near the same orthologous gene in more than one species. In many of the classical epimutants, epigenetic regulation of nearby transposon insertions can impact neighboring genes and cause phenotypic variation [21,24,25,26]. This additional regulation is in some cases beneficial; for example, for genes specifically expressed in the pollen [41,61]. The data presented here demonstrates that these events are both rare and likely lineage-specific. It is possible that the reported cases of differential methylation as a regulator of transcription are short-term innovations that are eventually replaced by genetically encoded regulation.

DNA methylation from an epimutational perspective

The mode of inheritance of symmetrically methylated cytosines motivates the interpretation of DNA methylation as a molecular modification that increases the complexity of the genetic code. While mutational processes affecting DNA sequence are well described, epimutational processes are poorly understood. DNA mutations rarely revert and occur in a largely random fashion throughout the genome [62]. In contrast, recent studies have shown that the transgenerational stability of DNA methylation is very context dependent [39,40]. Over short evolutionary times, epimutations are more likely to occur in euchromatic sequences and are biased away from heavily methylated repetitive sequences [39,40].

Over the longer evolutionary times examined here, we find that changes in genome content and structure are the major contributors to DNA methylation variation. While the majority of single site and regional methylation is found in repetitive sequences that are unlikely under evolutionary constraint, the remaining observed patterns in euchromatic sequence reflect lineage-specific evolution of transposons. This is particularly obvious in *A. lyrata*, which has experienced a recent invasion of transposable elements into euchromatic sequences [46] and subsequent elevation in the methylation rate of euchromatic features, particularly introns.

Large-scale structural changes that have perturbed the genome-wide DNA methylation landscape have also occurred in *A. thaliana* [48,49]. Loss of three repeat-rich centromeres in *A. thaliana* caused a decrease in DNA methylation in sequences flanking the ancestral centromeres. The impact of lineage-specific transposon evolution and subsequent methylation is similarly

evident in genic sequences. Approximately 40% of methylation in conserved exon sequence is species-specific. These sites are non-uniformly distributed near the 5' or 3' edges of genes, likely due to spreading from adjacent transposons [56,57,58]. These observations support the hypothesis that surveillance of transposons is the primary contributor to the genomic distribution of DNA methylation in plants. Since transposon content and genome structure vary extensively even over short evolutionary time periods, DNA methylation appears to be similarly variable. This is supported by the poor resolution of species relationships in a principal component analysis of DNA methylation and a nearly ten-fold increase in divergence between *A. lyrata* and *A. thaliana* when comparing DNA methylation as opposed to nucleotide sequence [46]. Together, these results indicate that DNA methylation as a non-canonical nucleotide is very rarely conserved over intermediate evolutionary times scales.

Despite the fact that we can estimate the epimutation rate of methylated cytosines and other parameters related to nucleotide mutations, it is misleading to equate DNA methylation changes to nucleotide substitutions. Our results indicate that the rapid evolution of repeat sequences is the major contributor to the equally rapid changes in the genomic distribution of DNA methylation. In this respect, it is more reasonable to regard DNA methylation primarily as a molecular phenotype resulting from the underlying genetic sequences. Although a few "pure" epialleles have been identified in nature, the majority of natural epimutations are linked to nearby transposon insertions or other genetic changes [21,24,25,26]. Fast evolution of repeat-sequences can, however, provide opportunities for lineage-specific cooption of DNA methylation for regulation of endogenous genes in response to various stimuli.

Materials and Methods

Experimental design

Seeds from the reference strain for each species (*A. thaliana* Col-0, *A. lyrata* MN47, *C. rubella* MTE) were sterilized with a 15 minute treatment of 30% bleach and 0.1% Triton X-100. Sterilized seeds were plated onto 0.5 × MS 0.7% agar plates with 1% sucrose. Each plate represented a single replicate consisting of 20 seedlings. In total, 7 replicates were sown and randomized into a 3 × 2 × 2 factorial design. The three factors in this experiment were species, tissue, and cold treatment. After sowing, plates were stratified in the dark at 4°C for 8 days, before being shifted to 23°C short-day conditions (8 hr light:16 hr dark). Plates were oriented vertically. After 6 days in 23°C, half of the plates were exposed to 4°C short-day conditions for 23 hours. At the end of the cold treatment, both control (23°C) and treated (4°C) samples were harvested. Root and shoot tissues were harvested independently. Plants were cut just above and below the root-shoot junction to separate the tissues and avoid cross contamination of tissue types. To minimize daily collection times, replicates were blocked by day.

RNA extraction and RNA-seq library preparation

Total RNA was isolated from three replicates of each factor combination using the Qiagen RNeasy Plant Mini Kit (catalog # 74904). An on-column DNase digestion was included (catalog # 79254). Total RNA integrity was confirmed on the Agilent BioAnalyzer. Illumina TruSeq RNA libraries were constructed using 3 µg of total RNA. Samples were randomized before library construction. The manufacturer's protocol was followed with one exception - 12 PCR cycles were used instead of the recommended 15. Libraries were quantified on an Agilent BioAnalyzer (DNA 1000 chip). Samples were normalized to 10 nM library molecules

and then pooled for sequencing. Three pools were constructed, each consisting of 12 random samples. Each pool was sequenced across three lanes of an Illumina GAII flowcell.

DNA extraction and bisulfite library preparation

DNA was extracted from two replicates of each factor combination using the Qiagen DNeasy Plant Mini Kit (catalog # 69104). DNA was quantified using the Qubit BR assay (Life Technologies, catalog # Q32853). Bisulfite libraries were constructed using modifications to the Illumina TruSeq DNA kit and published bisulfite library protocols [15,39]. Depending on the sample, starting material ranged from 200 ng to 1 µg. Changes to the manufacturer's protocol will be noted here. After shearing of genomic DNA with a Covaris S220 instrument, sheared lambda DNA was spiked into each sample (1:0.001 sample:lambda ratio) as a control, for accurate estimation of failure to bisulfite convert non-methylated cytosines. Samples were randomized before library construction. During the ligation step, the amount of adapter was adjusted based on the amount of starting material in each sample. For 1 µg of input DNA, 2.5 µl of adapter were used. Adapter input was scaled linearly for samples with less starting DNA. For the second AMPure bead clean up after the ligation step, the ratio of sample to beads was adjusted to 1:0.74. A final elution volume of 42.5 µl was used for this step. After ligation, 40 µl of eluate was transferred to a new tube for subsequent bisulfite treatment.

The Qiagen Epiect Plus Kit (catalog # 59124) was used for bisulfite treatment. The manufacturer's protocol for 'low concentrated and fragmented samples' was followed, using 85 µl of bisulfite mix for conversion. Clean up of the bisulfite reaction included ethanol as a final wash step. The sample was eluted in 17 µl. After bisulfite treatment samples were amplified using Pfu Cx HotStart Polymerase from Agilent (catalog # 600410) instead of the supplied PCR mix. Reaction conditions are all follows: 32.9 µl of water, 5 µl of 10× Pfu Cx Buffer, 5 µl of 2 mM dNTP, 1.6 µl of Illumina PCR Primer Cocktail, 0.5 µl of Cx Polymerase (2.5 U/µl), 5 µl of bisulfite-treated DNA eluate. Three PCR reactions were pooled for each bisulfite-treated sample. The following cycling conditions were used: 98°C - 30 seconds; 18 cycles of 98°C - 10 seconds, 65°C - 30 seconds, 72°C - 30 seconds; 72°C - 5 minutes. An AMPure bead clean up was used to purify the final PCR product (1:1 sample to bead ratio). Samples were eluted in 32.5 µl of Illumina supplied Resuspension buffer. 30 µl of the final eluate was transferred to a new plate for subsequent quantification and sequencing. Libraries were quantified using the Agilent BioAnalyzer (DNA 1000 chip). Libraries were diluted to 10 nM and then pooled. Samples were pooled based on genome size - and each pool consists of 2 random samples from each species. Four pools were constructed and each was sequenced across three lanes of the Illumina HiSeq 2000.

Bisulfite sequencing

We sequenced bisulfite-converted libraries with 2×101 base pair paired-end reads on an Illumina HiSeq 2000 instrument with conventional *A. thaliana* DNA genomic libraries in control lanes. Each sample contained 0.1% lambda DNA as an unmethylated control. We pooled six different samples in each lane. The Illumina RTA software (version 1.13.48) performed image analysis and base calling.

Processing and alignment of bisulfite-treated reads

Reads were filtered and trimmed as previously described [39]. Subsequently, trimmed reads were mapped against the corresponding reference genomes (Crubella_183, Alyrata_107,

Athaliana_167 (TAIR9) [46,47,50,51]. The lambda genome sequence was appended to each species genome sequence in order to estimate the false methylation rates of each sample. All reads were aligned using the mapping tool bismark v0.7.3 [63]. Applying the 'scoring matrix approach' of SHORE as previously described [39], we retrieved unique and non-duplicated read counts per position. Read and alignment statistics can be found in Table S2. All command line arguments are listed in Text S1. Raw reads are deposited at the European Nucleotide Archive under accession number PRJEB6701.

Determination of methylated sites

We used published methods [39], with a few exceptions. Here we retrieved incomplete bisulfite conversion rates, or false methylation rates (FMRs), from the alignments against the lambda genome rather than the chloroplast sequence. False methylation rates are found in Table S3. In addition, we combined the read counts of replicate samples after removing sites that were differentially methylated between replicates. The methylation rates for combined replicates were used for all subsequent analyses. The number of DMPs detected between replicates can be found in Table S19. In each species we required a methylation rate of at least 20% in one of the four tissue-treatment combinations in order for a site to be considered significantly methylated.

Identification of differentially methylated positions (DMPs)

To identify DMPs we followed published methods [39], but we required positions to have a methylation rate of at least 20% in one of the treatment combinations before performing Fisher's exact test. This increased statistical power by reducing the number of multiple testing corrections. Pairwise tests were not performed between all treatment combinations, instead only relevant comparisons were performed within each species (Root-23°C vs Shoot-23°C, Root-4°C vs Shoot-4°C, Root-23°C vs Root-4°C, Shoot-23°C vs Shoot-4°C).

Identification of methylated regions (MRs)

To detect contiguously methylated parts of the genome we modified a Hidden Markov Model (HMM) implementation [64]. Briefly, each cytosine can be in either an unmethylated or methylated state. The model trains methylation rate distributions for each state and sequence context (CG, CHG, CHH) independently using genome-wide data. In addition, transition probabilities between the states are trained. To make the original HMM implementation applicable to plant data, three different (beta binomial) distributions were estimated for each state (methylated and unmethylated) instead of just the single distribution used in mammals, which have almost only CG methylation [64]. To prevent identification of regions over uncovered bases, the genome was split at locations that lacked a covered cytosine position for 50 adjacent base pairs. On each of these segments, the most probable path through the methylation states was estimated after genome-wide parameter training. Transitions between states demarcated the methylated regions (MR). Replicates of each treatment combination were combined for this analysis. The combined read counts at cytosines were used to calculate methylation rates, train the HMM, and identify methylated regions. As a result, there is a single segmentation of the genome per treatment combination. Methylated regions were trimmed on both 5' and 3' ends by removing positions with a methylation rate

below 10%. Further details will be described in a manuscript by Hagmann, Becker et al. [65].

Identification of differentially methylated regions (DMRs)

Based on the MRs identified for each sample using the HMM algorithm described above, we selected regions of variable methylation state between samples to test for differential methylation. Due to the very large number of MRs, it was critical to reduce the number of tests performed to identify DMRs. By filtering MRs using the criteria outlined in a forthcoming manuscript by Hagmann, Becker et al. [65], we reduced the number of MRs four fold in each species. For each identified region, pairwise statistical tests were performed for the relevant comparisons listed above. The statistical test approximates the context-specific beta binomial distribution for the region of interest. Individual and joint distributions are approximated for two samples being compared. The statistical test compares the individual sample distributions to the joint distribution using a log-odds ratio. This ratio is compared against a chi-squared distribution to obtain confidence values. For each identified region, samples were assigned to groups by separating the samples with statistically significant methylation. To confirm groupings, we first combined read counts from treatment combinations in the same group. With the combined data, the same statistical test as described above was performed to test for differential methylation. Groups were confirmed in this way to identify and filter potentially erroneous DMRs. After false discovery rate (FDR) correction using Storey's method [66], regions with an FDR below 0.01 were defined as differentially methylated regions (DMRs). To resolve overlapping DMRs, we retained the non-overlapping regions containing the maximum number of samples with statistically significant differential methylation. Apart from the criterion used to resolve overlapping DMRs, the methods follow those that will be described in detail in a manuscript by Hagmann, Becker et al. [65].

Site-level conservation of methylation

We identified conserved sites using a published three-way whole genome alignment [47]. For CG sites, identical context was required while substitutions at the H positions were allowed in degenerate contexts as long as they did not mutate to G. Sites that transitioned contexts were not considered. Methylation rates for significantly methylated sites were then extracted from each species, tissue, and treatment combination for subsequent analysis.

Identification of 1:1:1 orthologous gene pairs

Three-way orthologs were identified using the reciprocal-best blast hit approach as implemented in the multiParanoid pipeline (inParanoid v. 4.1, blast v. 2.2.26) [67].

RNA sequencing

We sequenced each RNAseq library with 101 base pair single-end reads on the Illumina GAII instrument. We pooled twelve different samples in each lane. Each pool was sequenced over three lanes. The Illumina RTA software (version 1.13.48) performed image analysis and base calling.

Processing and alignment of RNAseq reads

Reads were trimmed using the shore import function in SHORE version 0.9.0 [68]. Command line arguments can be found in Text S1. This function simultaneously trims reads and separates samples by barcode. Since all samples were sequenced over three lanes, after lanes are de-multiplexed sample reads were

combined. Due to variable annotation qualities between species, only sequences annotated as CDS annotations were used to map RNA-seq reads. The following representative gene model annotation versions were used for each species: *Crubella_183*, *Alyrata_107*, *Athaliana_167* (TAIR10) [46,47,50,51]. Reads were aligned with one allowed mismatch to the appropriate annotation using bwa version 0.6.1 [69]. Read counts were obtained for each gene using a custom perl script. In summary, the script identified uniquely aligned read with a mapping quality score above 30 and stored the total read count for each target sequence. Read and alignment statistics can be found in Table S20. Raw reads are deposited at the European Nucleotide Archive under accession number PRJEB6701.

Differential expression analysis

Differentially expressed genes were identified using the R package edgeR (3.4.2) with minor modifications [70]. Using edgeR, we estimated the dispersion parameter for each gene using estimateGLMtagwiseDisp(). Next, we fit a negative binomial generalized linear model (GLM) using glmFit(). Significance testing for differential expression was performed using a custom GLM. Significance testing in edgeR was done via term-dropping of each factor level (likelihood ratio test), and as a result performed more statistical tests than necessary. To minimize multiple testing problems, we implemented a negative binomial GLM that tested for differential expression significance using an ANOVA [71]. Dispersion estimates from edgeR were provided to the modified GLM. Using this model, differential expression analysis was performed in two ways. First, expression analysis was performed within species. There were 12 samples consisting of three replicates and four unique treatment combinations. All representative gene models were considered. The following custom GLM model was used: expression~tissue*treatment. This included the main effects of tissue and treatment as well as their interaction. Secondly, we performed differential expression analysis between all species simultaneously. In this case, there are a total of 36 samples consisting of three replicates of each species, tissue, and treatment combination. Only 1:1:1 orthologous gene pairs were considered (14,395 in total). The following custom GLM model was used: expression~species*tissue*treatment. This includes the main effects of species, tissue, and treatment as well as all two and three-way interactions. Corrections for gene length were performed, but this did not impact the results and was subsequently ignored.

Repeat annotations

Transposon and repeat annotations for all three species were derived from the *Capsella rubella* genome paper [47,72,73].

Supporting Information

Figure S1 Effects of intron insertions of transposons on gene expression. Genes with and without TEs in their introns are compared. A gene is considered expressed if it had at least 3 RPKM in three of the twelve species-specific RNA-seq samples. (EPS)

Figure S2 Methylation rates of sequences flanking intron-inserted transposons. All cytosines in sequences flanking TEs in introns were extracted (+/-500 bp). Methylation rate for each annotated feature and context is calculated as the number of methylated cytosines over the total number of possible cytosines. Methylation rates are normalized to genome-wide methylation rates for each feature-context combination. Sites considered in our current analysis (intronic TE and +/-500 bp) were excluded from

the calculation of background methylation rates. This plot also accounts for expression of the gene containing the intronic TE. (EPS)

Figure S3 Methylation rates at genomic features of expressed genes. Genome average of methylation rates for each genomic feature. Similar to figure 2B, except annotations are only considered for genes that are expressed. A gene is considered expressed if it received at least 3 RPKM in three of the twelve species-specific RNA-seq samples. Methylation rates are normalized to the outgroup species *C. rubella*. (EPS)

Figure S4 Genomic distribution of MRs and DMRs. A) Circos plots [74] to demonstrate the genomic distribution of MRs and DMRs in *C. rubella*, *A. lyrata*, and *A. thaliana*. Chromosome number is indicated on the inner circle. Data is plotted for 500 kb windows. (EPS)

Figure S5 Relationship between gene body methylation and gene expression. Gene body methylation rates are plotted against either gene expression (\log_2) deciles or coefficient of variation (CV) deciles. When comparing gene body methylation with gene expression the Spearman rank correlation coefficient in *C. rubella* = 0.21, *A. lyrata* = 0.23, and *A. thaliana* = 0.24. In contrast, when comparing gene body methylation with CV the Spearman rank correlation coefficient in *C. rubella* = -0.34, *A. lyrata* = -0.19, and *A. thaliana* = -0.33. (EPS)

Figure S6 Annotation of methylated site classes in three-way alignments. Feature annotation is shown for each methylation context. Site classes are as follows: Aligned - all C in three-way alignments. mC - methylated sites within a species. Consv. (3 species) - sites that are methylated in all three species. Gain - sites that are methylated in a single species. Loss - sites that have lost methylation in a single species. (EPS)

Figure S7 Transposon categories for aligned methylated site classes. The top 5% of windows (10 kb) for three-way conserved sites, gains, and losses were identified. As a control, an equal number of random genomic windows were chosen. Shown is the number of bases annotated as a transposon category for the top 5% of windows in each site class normalized to the control annotation. (EPS)

Figure S8 Centromere loss is not associated with methylation loss at aligned cytosines. Fraction of species-specific losses in methylation is plotted for each ortholog residing within ancestral centromere boundaries. Orthologs were categorized based on genomic position, either in or outside of ancestral centromere boundaries. Centromere boundaries were defined in *C. rubella* using repeat density (Fig. 3, 0.3 threshold). Orthologs residing in maintained ancestral centromeres ("No Loss") were compared to orthologs residing in ancestral centromeres lost in *A. thaliana* ("Loss"). (EPS)

Figure S9 Conserved methylated sites associated with conservation of exon length. Fraction of site categories that reside in exons with conserved lengths across all three species or exons of variable lengths. (EPS)

Figure S10 Distribution of cytosines across exons. The density of exon methylation at aligned cytosines is shown for conserved methylated sites as well as for lineage-specific gains and losses of methylation. On top is the density of non-methylated aligned cytosines. There is no bias in location within an exon for non-methylated sites. (EPS)

Figure S11 Differential gene expression. For each model, within species (top) and between species (bottom), the number of differentially expressed genes (absolute and as a fraction of expressed genes) is shown for each main effect and all interactions ($p < 0.05$). (EPS)

Figure S12 Species mC relationship of replicates. A) Principal component analysis on mC rates at aligned methylated positions. All contexts are considered (see Fig. 6B,C and Table S10 for further description of mC sites). Unlike figure 7, this plot considers the mC rate of each replicate at all aligned methylated positions. (EPS)

Figure S13 Genomic distribution of DMPs. A) Circos plots [74] to demonstrate the genomic distribution of DMPs in *C. rubella*, *A. lyrata*, and *A. thaliana*. Plots are separate for tissue specific DMPs (root and shoot) or treatment specific DMPs (23°C and 4°C). Chromosome number is indicated on the inner circle. Data is plotted for 500 kb windows. (EPS)

Figure S14 Conservation of DMRs in the absence of sequence alignments. The total number of orthologous genes containing a DMR in one, two, or three species is shown. Location of DMR overlap is separated by genomic feature. Upstream region is considered 1 kb before the start codon. Asterisk indicates two or three-way sharing of DMRs that exceeds permutation values. (EPS)

Table S1 References for genome size. References for the genome size (in pg and Mb) as well as the total size of the genome assembly are listed for each species. Genome size references are derived from the Kew Royal Botanic Gardens Plant DNA C-values database. (XLSX)

Table S2 Bisulfite-sequencing coverage and alignment statistics. For each sample, the total number of sequenced reads (paired and single) is shown. Also, the total number of CG, CHG, and CHH sites covered is reported along with the average genome-wide coverage of each context. (XLSX)

Table S3 False methylation rates by coverage bin. The incomplete bisulfite conversion rate, or false methylation rate (FMR), for each sample is shown by coverage bin. For each bin, FMR is calculated as the number of cytosines in lambda DNA that are not converted to U (T in the DNA sequence) after bisulfite treatment over the total number of converted (U/T) and unconverted (C) reads. (XLSX)

Table S4 MR and DMR statistics by sample (A) and species (B). Mean and median length of region, total number of regions, and genomic bases covered by regions are shown. Sample statistics were calculated from the combination of biological replicates. (XLSX)

Table S5 Genome alignment metrics. Number of bases covered in three-way whole genome alignments is shown. In addition, total number of bases in methylated site classes is shown. (XLSX)

Table S6 MR presence at genomic features. The number of genes in each species annotation that contain a MR is shown. Upstream refers to 1 kb upstream of the start codon. The number of orthologous genes with an overlapping MR is also shown. (XLSX)

Table S7 P values of pairwise MR overlap. To test the significance of MR co-occurrence at orthologous genes a hypergeometric test was used. Significance of each test is shown here. (XLSX)

Table S8 Two and three-way species MR overlap. The number of orthologs that contain an MR in one, two, or three species is shown (A). Permutation analysis was performed to estimate the random occurrence of one, two, and three-way overlap (10,000 permutation tests). Maximum permutation values are shown in (B). Features where the data exceeds the maximum permutation value are indicated in (C). (XLSX)

Table S9 Gene body methylation by context. The total numbers of genes with CG, CHG, or CHH gene body methylation are shown for all genes (A) and orthologous genes (B). (XLSX)

Table S10 Three-way genome alignment site classes by context. Total numbers of CG, CHG, and CHH sites for each alignment site class are shown. (XLSX)

Table S11 DMP statistics by comparison. Total numbers of DMPs in each tissue and treatment comparison are shown. (XLSX)

Table S12 DMR statistics by comparison. Total numbers of DMRs in each tissue and treatment comparison are shown. (XLSX)

Table S13 DMR presence at genomic features. The number of genes in each species' annotation that contain a DMR is shown. Upstream refers to 1 kb upstream of the start codon. The number of orthologous genes with an overlapping DMR is also shown. (XLSX)

Table S14 P values of pairwise DMR overlap. To test the significance of DMR co-occurrence at orthologous genes a hypergeometric test was used. Significance of each test is shown here. (XLSX)

Table S15 Two and three-way species DMR overlap. The number of orthologs that contain a DMR in one, two, or three species is shown (A). Permutation analysis was performed to estimate the random occurrence of one, two, and three-way overlap (10,000 permutation tests). Maximum permutation values

are shown in (B). Features where the data exceeds the maximum permutation value are indicated in (C). (XLSX)

Table S16 DMPs at genomic features. The number of genes in each species annotation that contain a DMP is shown. Upstream refers to 1 kb upstream of the start codon. The number of orthologous genes with an overlapping DMP is also shown. (XLSX)

Table S17 DMP correlation with gene expression by feature. Spearman rank correlation coefficient was calculated between the direction of differential methylation and the appropriate \log_2 fold change. Correlation coefficients were calculated separately for tissue and treatment specific DMPs. An NA value indicates that there were too few genes in a given category. Expression values are from the intraspecific expression analysis. (XLSX)

Table S18 DMR correlation with gene expression by feature. Spearman rank correlation coefficients when comparing the degree of differential methylation for each context (extracted from the HMM model) with the appropriate \log_2 fold change. All annotated genes overlapping a DMR were considered. Expression values are from the intraspecific expression analysis. Correlation coefficients were calculated only for tissue-specific DMRs as there are too few treatment-specific DMRs. Results are only shown for DMRs overlapping CDS, intron, and upstream sequences because too few expressed genes reside in the other categories (5' and 3' UTRs). Upstream refers to 1 kb upstream of the start codon. (XLSX)

Table S19 Number of DMPs between replicates. For each species, tissue, treatment combination, differentially methylated positions were identified between biological replicates. The total number of DMPs for each comparison is listed. These positions were removed from all further analyses. (XLSX)

Table S20 RNA-seq sequencing coverage and alignment statistics. For each sample, the total number of RNA sequencing reads is shown. Read counts are also shown for mapped reads, uniquely mapped reads, and the reads that passed a mapping quality threshold (30). (XLSX)

Text S1 Command lines for alignments. Command lines and arguments for the processing of bisulfite reads and RNA-seq reads. (TXT)

Acknowledgments

We thank Florian Maumus and Hadi Quesneville for early access to their transposon annotation of the three genomes.

Author Contributions

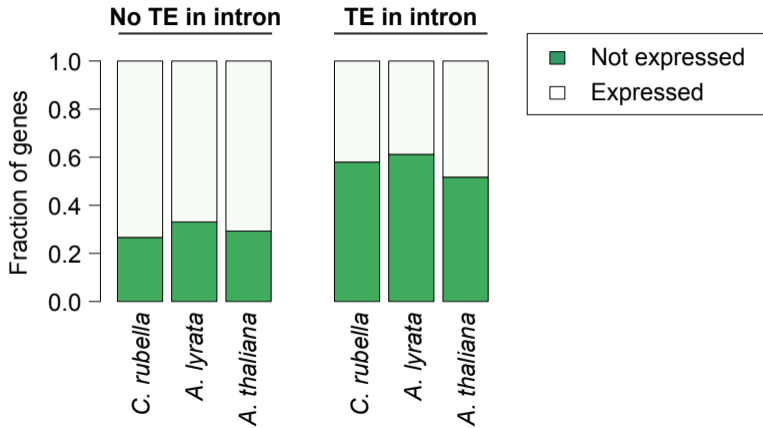
Conceived and designed the experiments: DKS DK CB DW. Performed the experiments: DKS DK. Analyzed the data: DKS DK JH CB. Contributed to the writing of the manuscript: DKS DK DW.

References

- Feng S, Cokus SJ, Zhang X, Chen PY, Bostick M, et al. (2010) Conservation and divergence of methylation patterning in plants and animals. *Proc Natl Acad Sci U S A* 107: 8689–8694.
- Zemach A, McDaniel IE, Silva P, Zilberman D (2010) Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science* 328: 916–919.
- Suzuki MM, Bird A (2008) DNA methylation landscapes: provocative insights from epigenomics. *Nat Rev Genet* 9: 465–476.
- Huff JT, Zilberman D (2014) Dnmt1-independent CG methylation contributes to nucleosome positioning in diverse eukaryotes. *Cell* 156: 1286–1297.
- Gruenbaum Y, Navehmany T, Cedar H, Razin A (1981) Sequence specificity of methylation in higher-plant DNA. *Nature* 292: 860–862.
- Zhang X, Yazaki J, Sundaresan A, Cokus S, Chan SW, et al. (2006) Genome-wide high-resolution mapping and functional analysis of DNA methylation in arabidopsis. *Cell* 126: 1189–1201.

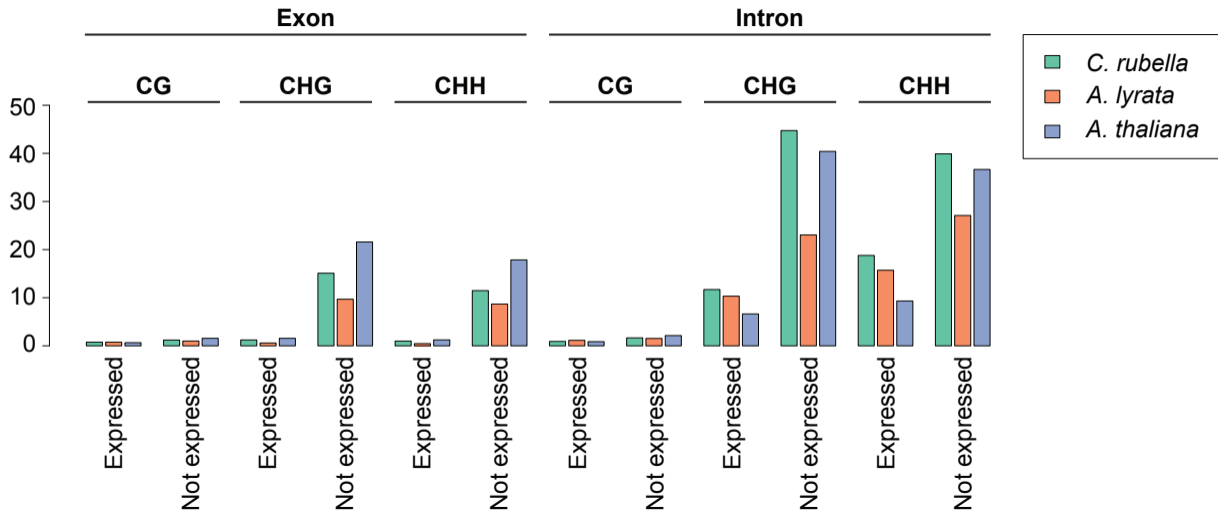
7. Zilberman D, Gehring M, Tran RK, Ballinger T, Henikoff S (2007) Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. *Nat Genet* 39: 61–69.
8. Rabinowicz PD, Schutz K, Dedhia N, Yordan C, Parnell LD, et al. (1999) Differential methylation of genes and retrotransposons facilitates shotgun sequencing of the maize genome. *Nat Genet* 23: 305–308.
9. Regulski M, Lu Z, Kendall J, Donoghue MT, Reinders J, et al. (2013) The maize methylome influences mRNA splice sites and reveals widespread paramutation-like switches guided by small RNA. *Genome Res* 23: 1651–1662.
10. Li X, Zhu J, Hu F, Ge S, Ye M, et al. (2012) Single-base resolution maps of cultivated and wild rice methylomes and regulatory roles of DNA methylation in plant gene expression. *BMC Genomics* 13: 300.
11. Takuno S, Gaut BS (2013) Gene body methylation is conserved between plant orthologs and is of evolutionary consequence. *Proc Natl Acad Sci U S A* 110: 1797–1802.
12. Slotkin RK, Martienssen R (2007) Transposable elements and the epigenetic regulation of the genome. *Nat Rev Genet* 8: 272–285.
13. Lisch D (2013) How important are transposons for plant evolution? *Nat Rev Genet* 14: 49–61.
14. Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, et al. (2008) Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature* 452: 215–219.
15. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, et al. (2008) Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell* 133: 523–536.
16. Lister R, Pelizzola M, Downen RH, Hawkins RD, Hon G, et al. (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 462: 315–322.
17. Coleman-Derr D, Zilberman D (2012) Deposition of histone variant H2A.Z within gene bodies regulates responsive genes. *PLoS Genet* 8: e1002988.
18. Schmitz RJ, Schultz MD, Urich MA, Nery JR, Pelizzola M, et al. (2013) Patterns of population epigenomic diversity. *Nature* 495: 193–198.
19. Widman N, Feng S, Jacobsen SE, Pellegrini M (2014) Epigenetic differences between shoots and roots in Arabidopsis reveals tissue-specific regulation. *Epigenetics* 9: 236–242.
20. Downen RH, Pelizzola M, Schmitz RJ, Lister R, Downen JM, et al. (2012) Widespread dynamic DNA methylation in response to biotic stress. *Proc Natl Acad Sci U S A* 109: E2183–2191.
21. Morgan HD, Sutherland HG, Martin DI, Whitelaw E (1999) Epigenetic inheritance at the agouti locus in the mouse. *Nat Genet* 23: 314–318.
22. Cubas P, Vincent C, Coen E (1999) An epigenetic mutation responsible for natural variation in floral symmetry. *Nature* 401: 157–161.
23. Manning K, Tor M, Poole M, Hong Y, Thompson AJ, et al. (2006) A naturally occurring epigenetic mutation in a gene encoding an SBP-box transcription factor inhibits tomato fruit ripening. *Nat Genet* 38: 948–952.
24. Martin A, Troade C, Boualem A, Rajab M, Fernandez R, et al. (2009) A transposon-induced epigenetic change leads to sex determination in melon. *Nature* 461: 1135–1138.
25. Das OP, Messing J (1994) Variegated phenotype and developmental methylation changes of a maize allele originating from epimutation. *Genetics* 136: 1121–1141.
26. Liu J, He Y, Amasino R, Chen X (2004) siRNAs targeting an intronic transposon in the regulation of natural flowering behavior in Arabidopsis. *Genes Dev* 18: 2873–2878.
27. Miura K, Agetsuma M, Kitano H, Yoshimura A, Matsuoka M, et al. (2009) A metastable DWARF1 epigenetic mutant affecting plant stature in rice. *Proc Natl Acad Sci U S A* 106: 11218–11223.
28. Heard E, Martienssen RA (2014) Transgenerational epigenetic inheritance: myths and mechanisms. *Cell* 157: 95–109.
29. Lindroth AM, Cao X, Jackson JP, Zilberman D, McCallum CM, et al. (2001) Requirement of CHROMOMETHYLASE3 for maintenance of CpXpG methylation. *Science* 292: 2077–2080.
30. Bestor T, Laudano A, Mattaliano R, Ingram V (1988) Cloning and sequencing of a cDNA encoding DNA methyltransferase of mouse cells. The carboxyl-terminal domain of the mammalian enzymes is related to bacterial restriction methyltransferases. *J Mol Biol* 203: 971–983.
31. Finnegan EJ, Dennis ES (1993) Isolation and identification by sequence homology of a putative cytosine methyltransferase from Arabidopsis thaliana. *Nucleic Acids Res* 21: 2383–2388.
32. Leonhardt H, Page AW, Weier HU, Bestor TH (1992) A targeting sequence directs DNA methyltransferase to sites of DNA replication in mammalian nuclei. *Cell* 71: 865–873.
33. Chuang LSH, Ian HI, Koh TW, Ng HH, Xu GL, et al. (1997) Human DNA (cytosine-5) methyltransferase PCNA complex as a target for p21(WAF1). *Science* 277: 1996–2000.
34. Pelissier T, Thalmeir S, Kempe D, Sanger HL, Wassenegger M (1999) Heavy de novo methylation at symmetrical and non-symmetrical sites is a hallmark of RNA-directed DNA methylation. *Nucleic Acids Res* 27: 1625–1634.
35. Cao X, Jacobsen SE (2002) Locus-specific control of asymmetric and CpNpG methylation by the DRM and CMT3 methyltransferase genes. *Proc Natl Acad Sci U S A* 99 Suppl 4: 16491–16498.
36. Law JA, Jacobsen SE (2010) Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat Rev Genet* 11: 204–220.
37. Chan SW, Zilberman D, Xie Z, Johansen LK, Carrington JC, et al. (2004) RNA silencing genes control de novo DNA methylation. *Science* 303: 1336.
38. Zemach A, Kim MY, Hsieh PH, Coleman-Derr D, Eshed-Williams L, et al. (2013) The Arabidopsis nucleosome remodeler DDM1 allows DNA methyltransferases to access H1-containing heterochromatin. *Cell* 153: 193–205.
39. Becker C, Hagemann J, Müller J, Koenig D, Stegle O, et al. (2011) Spontaneous epigenetic variation in the Arabidopsis thaliana methylome. *Nature* 480: 245–249.
40. Schmitz RJ, Schultz MD, Lewsey MG, O'Malley RC, Urich MA, et al. (2011) Transgenerational epigenetic instability is a source of novel methylation variants. *Science* 334: 369–373.
41. Calarco JP, Borges F, Donoghue MT, Van Ex F, Jullien PE, et al. (2012) Reprogramming of DNA methylation in pollen guides epigenetic inheritance via small RNA. *Cell* 151: 194–205.
42. Takuno S, Gaut BS (2012) Body-methylated genes in Arabidopsis thaliana are functionally important and evolve slowly. *Mol Biol Evol* 29: 219–227.
43. Long HK, Sims D, Heger A, Blackledge NP, Kutter C, et al. (2013) Epigenetic conservation at gene regulatory elements revealed by non-methylated DNA profiling in seven vertebrates. *eLife* 2: e00348.
44. Beilstein MA, Nagalingum NS, Clements MD, Manchester SR, Mathews S (2010) Dated molecular phylogenies indicate a Miocene origin for Arabidopsis thaliana. *Proc Natl Acad Sci U S A* 107: 18724–18728.
45. Johnston JS, Pepper AE, Hall AE, Chen ZJ, Hodnett G, et al. (2005) Evolution of genome size in Brassicaceae. *Ann Bot* 95: 229–235.
46. Hu TT, Pattyn P, Bakker EG, Cao J, Cheng JF, et al. (2011) The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nat Genet* 43: 476–481.
47. Slotte T, Hazzouri KM, Agren JA, Koenig D, Maumus F, et al. (2013) The Capsella rubella genome and the genomic consequences of rapid mating system evolution. *Nat Genet* 45: 831–835.
48. Yogeeswaran K, Frary A, York TL, Amenta A, Lesser AH, et al. (2005) Comparative genome analyses of Arabidopsis spp.: inferring chromosomal rearrangement events in the evolutionary history of A. thaliana. *Genome Res* 15: 505–515.
49. Lysak MA, Berr A, Pecinka A, Schmidt R, McBreen K, et al. (2006) Mechanisms of chromosome number reduction in Arabidopsis thaliana and related Brassicaceae species. *Proc Natl Acad Sci U S A* 103: 5224–5229.
50. The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* 408: 796–815.
51. Swarbreck D, Wilks C, Lamesch P, Berardini TZ, Garcia-Hernandez M, et al. (2008) The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res* 36: D1009–1014.
52. Lysak MA, Koch MA, Beaulieu JM, Meister A, Leitch IJ (2009) The dynamic ups and downs of genome size evolution in Brassicaceae. *Mol Biol Evol* 26: 85–98.
53. Bennett MD, Leitch IJ, Price HJ, Johnston JS (2003) Comparisons with Caenorhabditis (similar to 100 Mb) and Drosophila (similar to 175 Mb) using flow cytometry show genome size in Arabidopsis to be similar to 157 Mb and thus similar to 25% larger than the Arabidopsis genome initiative estimate of similar to 125 Mb. *Ann Bot* 91: 547–557.
54. Bennett MD, Leitch IJ (2011) Nuclear DNA amounts in angiosperms: targets, trends and tomorrow. *Ann Bot* 107: 467–590.
55. Schmitz RJ, He Y, Valdes-Lopez O, Khan SM, Joshi T, et al. (2013) Epigenome-wide inheritance of cytosine methylation variants in a recombinant inbred population. *Genome Res* 23: 1663–1674.
56. Arnaud P, Goubely C, Pelissier T, Deragon JM (2000) SINE retrotransposons can be used in vivo as nucleation centers for de novo methylation. *Mol Cell Biol* 20: 3434–3441.
57. Sun FL, Haynes K, Simpson CL, Lee SD, Collins L, et al. (2004) cis-Acting determinants of heterochromatin formation on Drosophila melanogaster chromosome four. *Mol Cell Biol* 24: 8210–8220.
58. Saze H, Kakutani T (2007) Heritable epigenetic mutation of a transposon-flanked Arabidopsis gene due to lack of the chromatin-remodeling factor DDM1. *EMBO J* 26: 3641–3652.
59. Chodavarapu RK, Feng S, Bernatavichute YV, Chen PY, Stroud H, et al. (2010) Relationship between nucleosome positioning and DNA methylation. *Nature* 466: 388–392.
60. Baubec T, Fink A, Scheid OM, Pecinka A (2014) Meristem-specific expression of epigenetic regulators safeguards transposon silencing in Arabidopsis. *EMBO Rep* 15: 446–452.
61. Slotkin RK, Vaughn M, Borges F, Tanurdzic M, Becker JD, et al. (2009) Epigenetic reprogramming and small RNA silencing of transposable elements in pollen. *Cell* 136: 461–472.
62. Lynch M (2007) The Origins of Genome Architecture. Sunderland, MA: Sinauer Associates. 389 p.
63. Krueger F, Andrews SR (2011) Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* 27: 1571–1572.
64. Molaro A, Hodges E, Fang F, Song Q, McCombie WR, et al. (2011) Sperm methylation profiles reveal features of epigenetic inheritance and evolution in primates. *Cell* 146: 1029–1041.
65. Hagemann J, Becker C, Müller J, Stegle O, Meyer RC, et al. (2014) Century-scale methylome stability in a recently diverged Arabidopsis thaliana lineage. *bioRxiv* doi: 10.1101/009225.

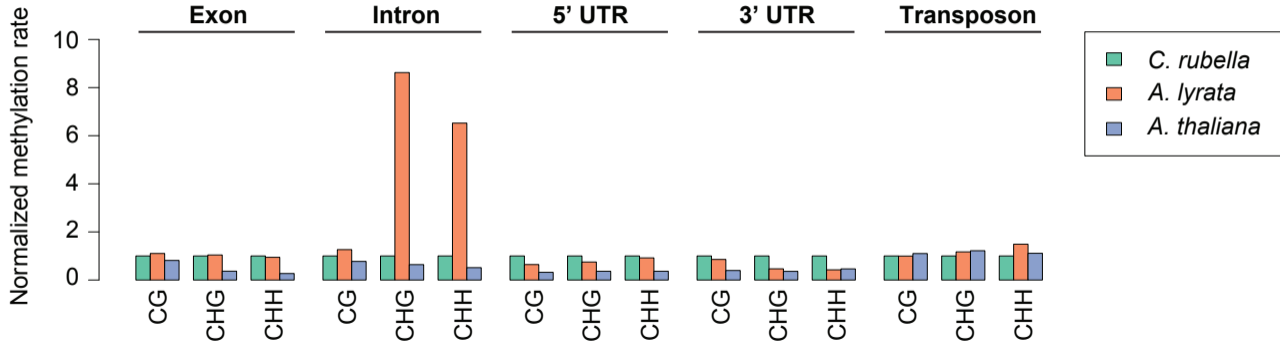
66. Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. *Proc Natl Acad Sci USA* 100: 9440–9445.
67. Remm M, Storm CE, Sonnhammer EL (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol* 314: 1041–1052.
68. Ossowski S, Schneeberger K, Clark RM, Lanz C, Warthmann N, et al. (2008) Sequencing of natural strains of *Arabidopsis thaliana* with short reads. *Genome Res* 18: 2024–2033.
69. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25: 1754–1760.
70. Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26: 139–140.
71. Koenig D, Jimenez-Gomez JM, Kimura S, Fulop D, Chitwood DH, et al. (2013) Comparative transcriptomics reveals patterns of selection in domesticated and wild tomato. *Proc Natl Acad Sci U S A* 110: E2655–2662.
72. Maumus F, Quesneville H (2014) Deep investigation of *Arabidopsis thaliana* junk DNA reveals a continuum between repetitive elements and genomic dark matter. *PLoS ONE* 9: e94101.
73. Maumus F, Quesneville H (2014) Ancestral repeats have shaped epigenome and genome composition for millions of years in *Arabidopsis thaliana*. *Nat Commun* 5: 4104.
74. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, et al. (2009) Circos: an information aesthetic for comparative genomics. *Genome Res* 19: 1639–1645.

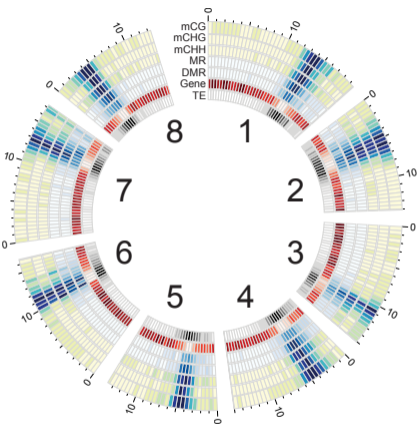
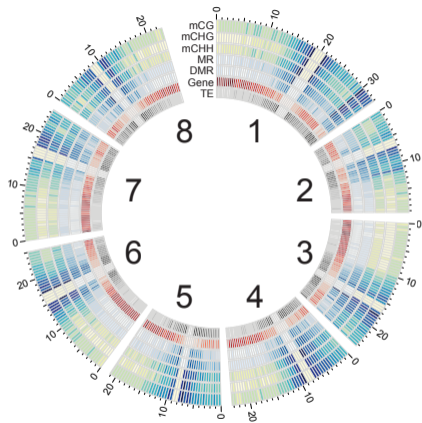
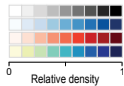
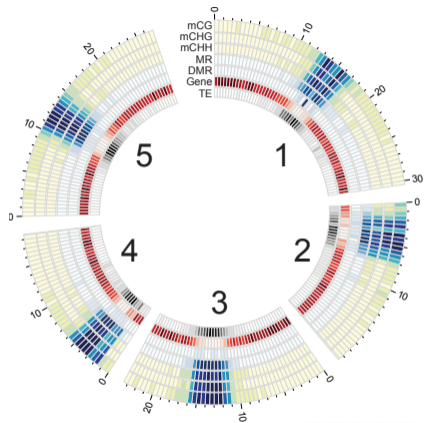
S1

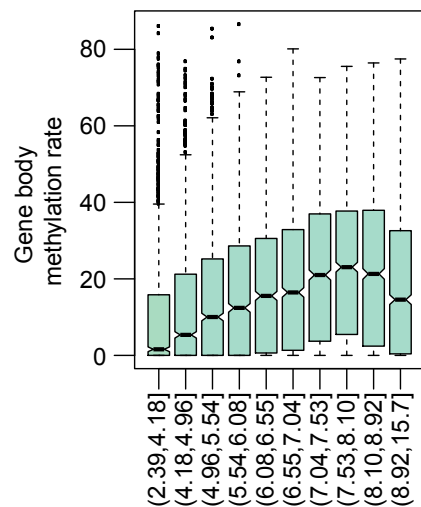
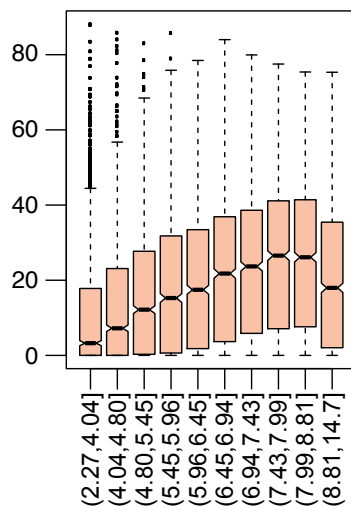
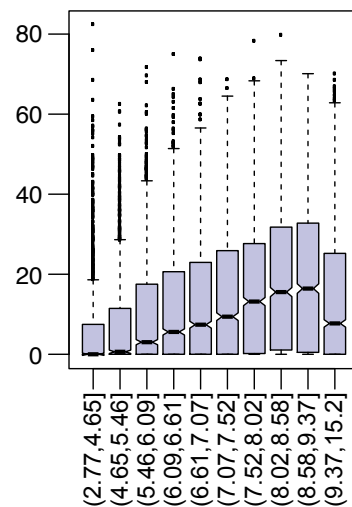
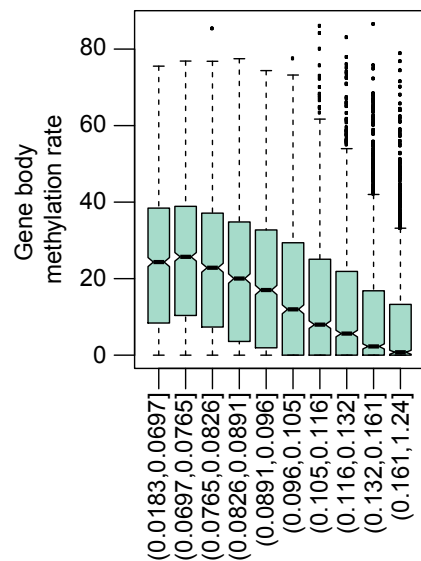
S2

Normalized methylation rate

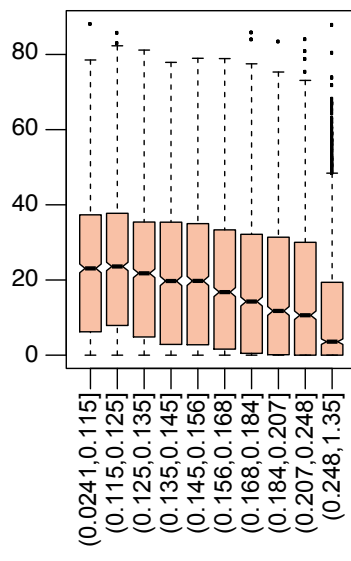


S3

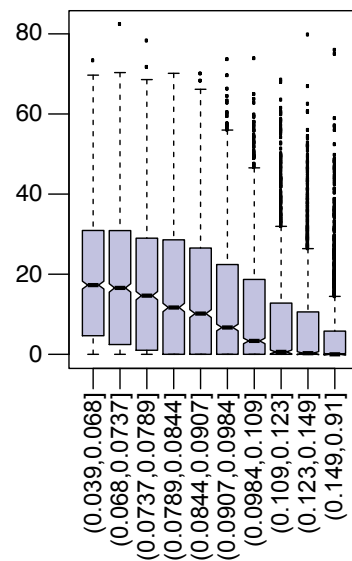
S4*C. rubella**A. lyrata**A. thaliana*

C. rubella \log_2 (expression) quantile*A. lyrata* \log_2 (expression) quantile*A. thaliana* \log_2 (expression) quantile*C. rubella*

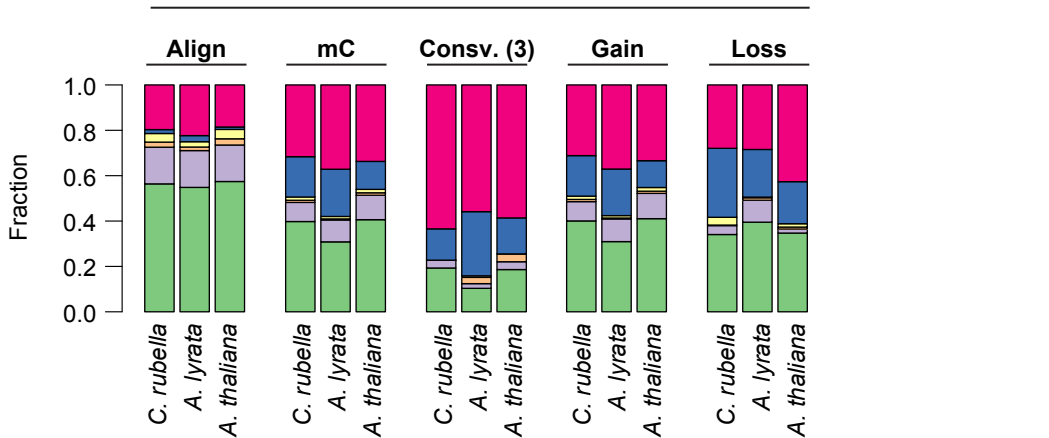
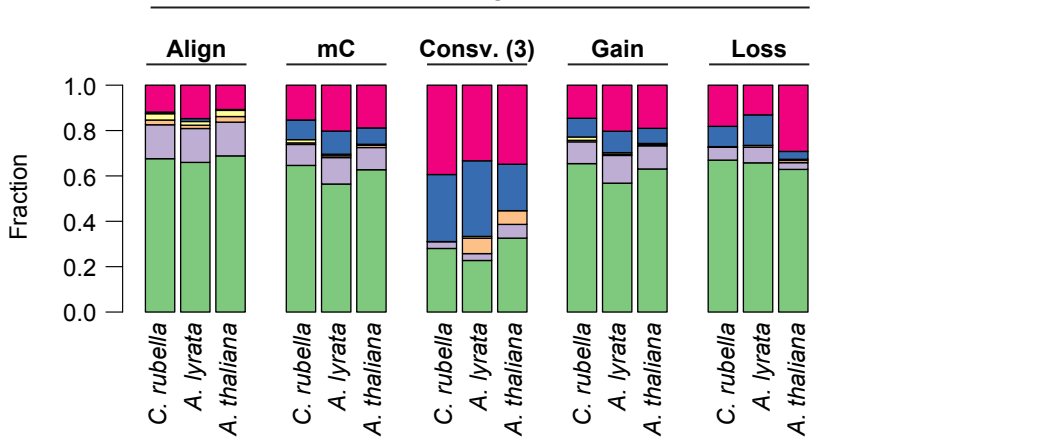
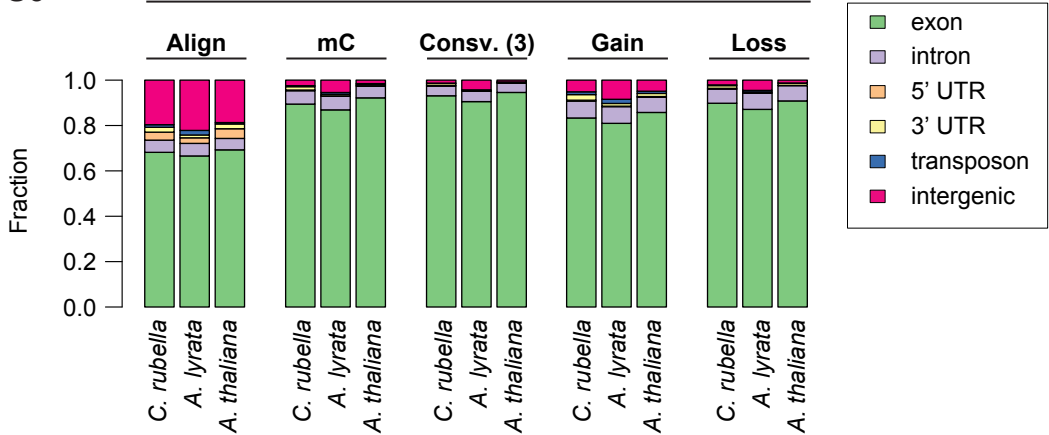
CV quantile

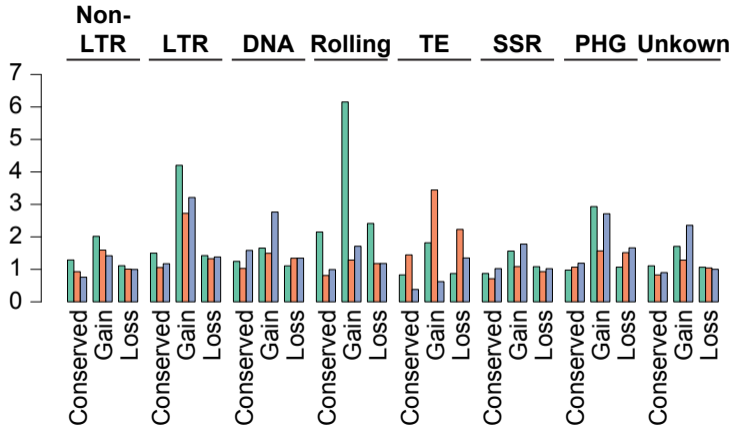
A. lyrata

CV quantile

A. thaliana

CV quantile



S7TE bases normalized
to control windows

C. rubella
A. lyrata
A. thaliana

S8

Loss/Aligned C

0.05
0.04
0.03
0.02
0.01
0.00

Loss

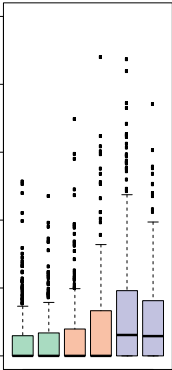
No Loss

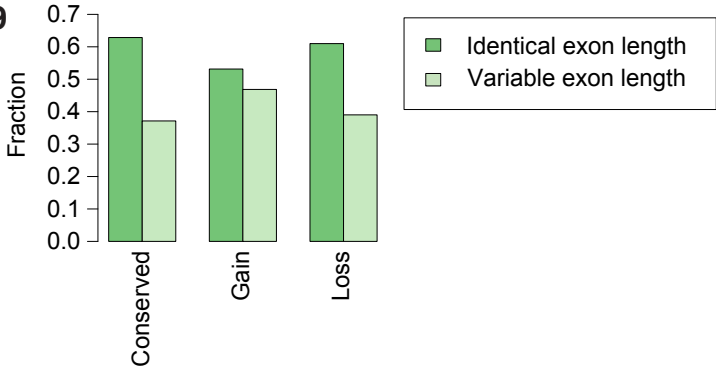
Loss

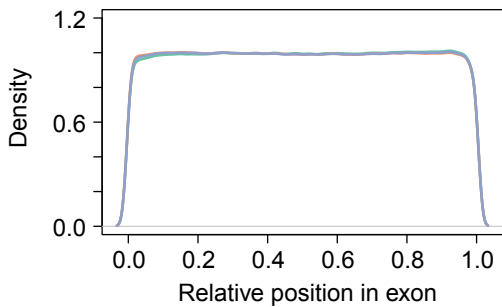
No Loss

Loss

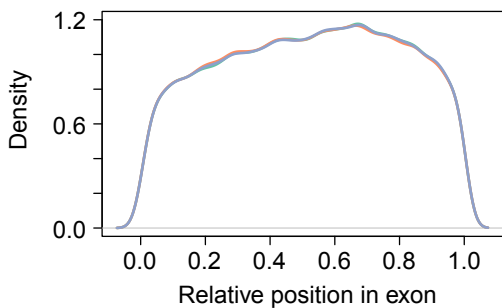
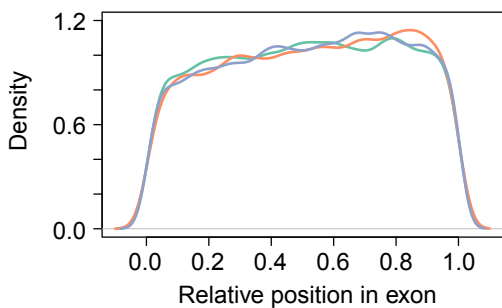
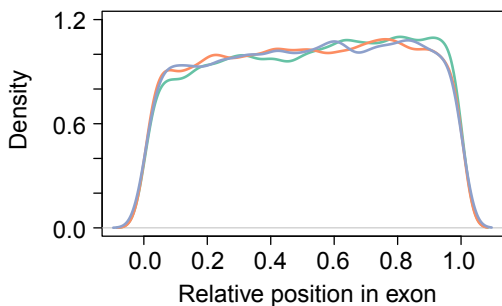
No Loss

*C. rubella**A. lyrata**A. thaliana*

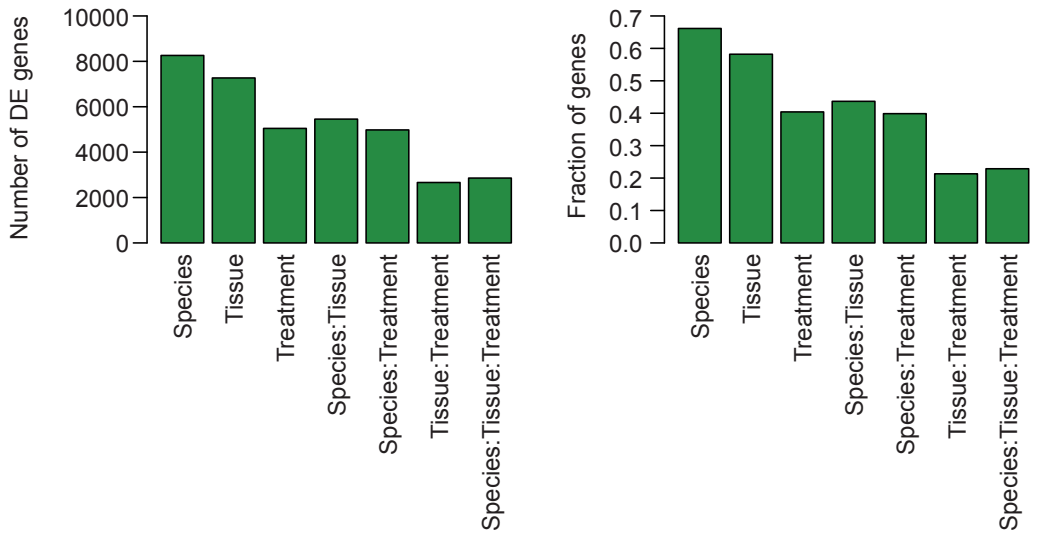
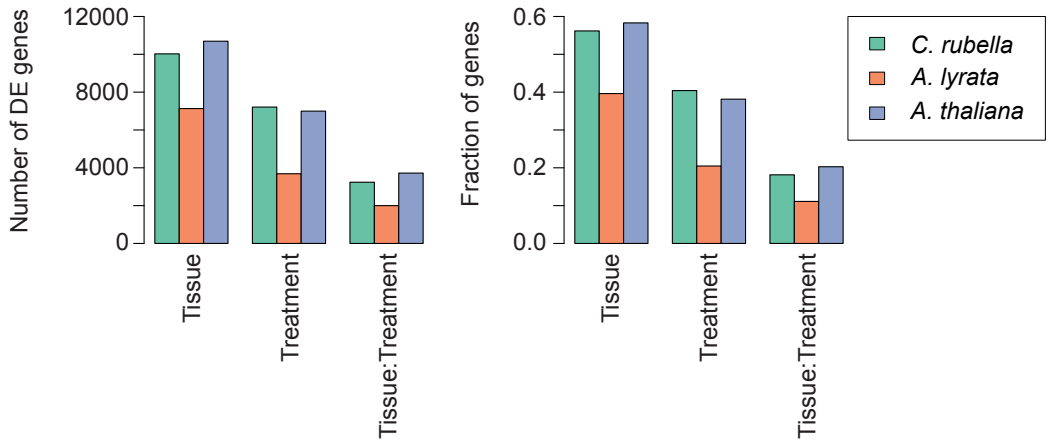
S9

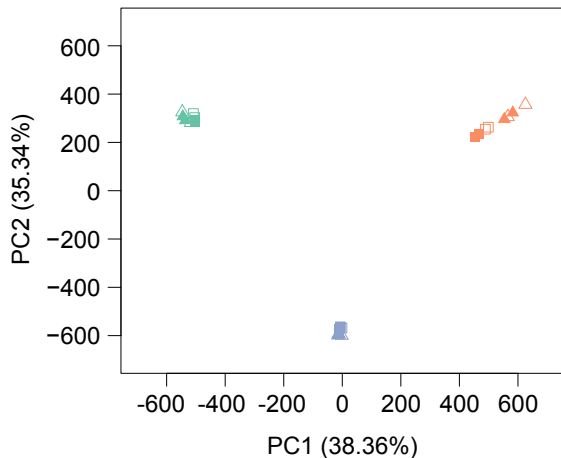
S10**Aligned non-mC**

C. rubella
A. lyrata
A. thaliana

Conserved-mC**Loss of mC****Gain of mC**

S11



S12

□ *C. rubella* - root - 23°C

△ *C. rubella* - shoot - 23°C

■ *C. rubella* - root - 4°C

▲ *C. rubella* - shoot - 4°C

□ *A. lyrata* - root - 23°C

△ *A. lyrata* - shoot - 23°C

■ *A. lyrata* - root - 4°C

▲ *A. lyrata* - shoot - 4°C

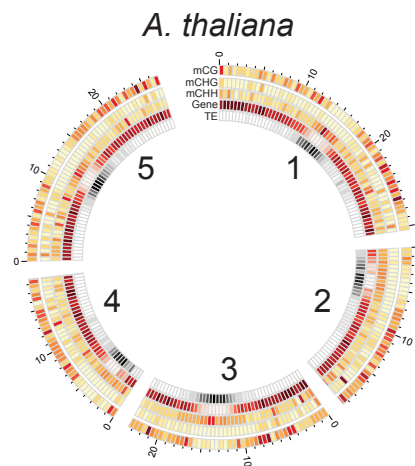
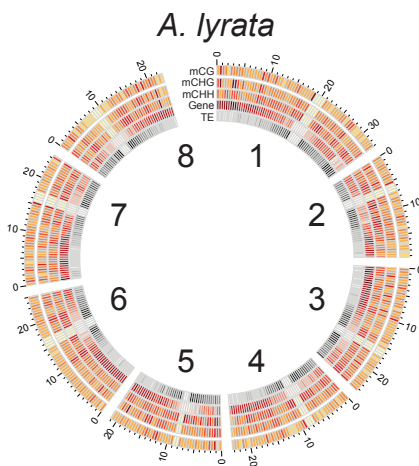
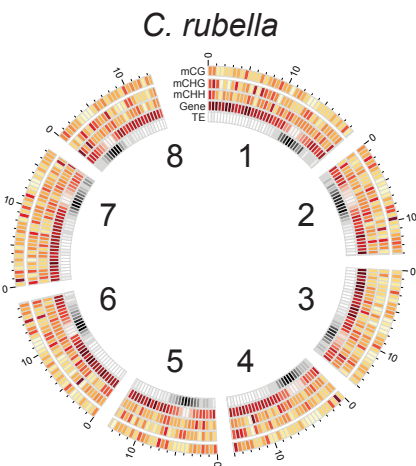
□ *A. thaliana* - root - 23°C

△ *A. thaliana* - shoot - 23°C

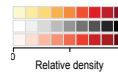
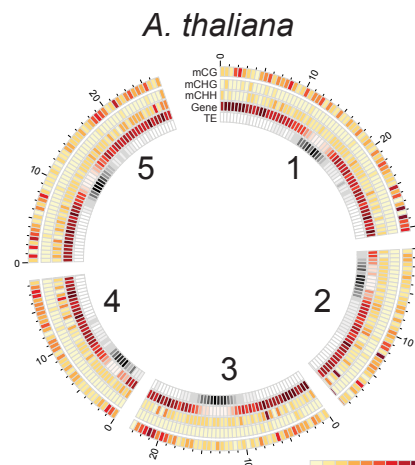
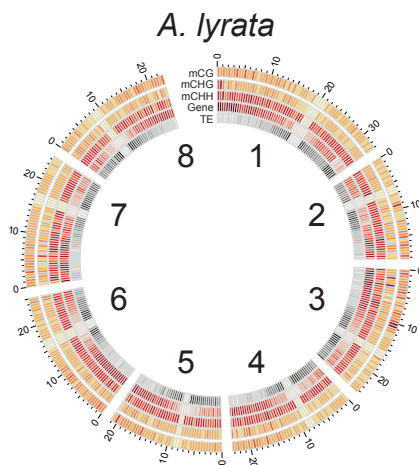
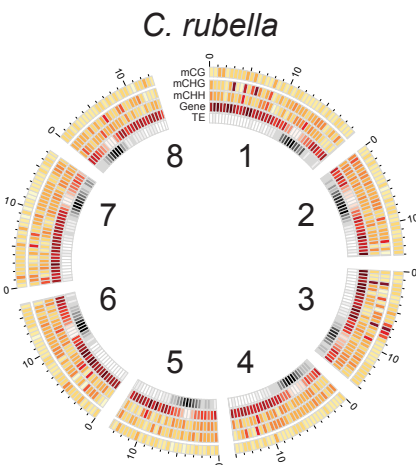
■ *A. thaliana* - root - 4°C

▲ *A. thaliana* - shoot - 4°C

Root-Shoot Comparisons



Control-Treatment Comparisons



S14

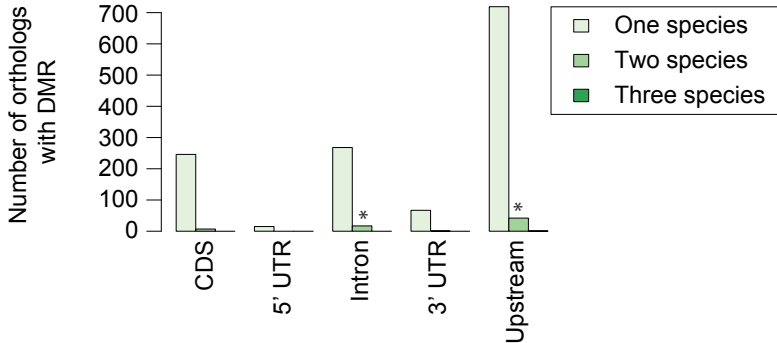


Table S1. References for genome size.

Species	Estimated size (pg)*	Estimated size (Mbp)*	Genome assembly	References for genome size*	References for genome assembly
<i>C. rubella</i>	0.22	215 (219)	135	Lysak et al., 2009; Bennett et al., 2011; Slotte et al., 2013	Slotte et al., 2013
<i>A. lyrata</i>	0.25	245	207	Lysak et al., 2009; Bennett et al., 2011	Hu et al., 2011
<i>A. thaliana</i>	0.16	156	119	Bennett et al., 2003; Bennett et al., 2005	Swarbreck et al., 2008

*From the Kew Royal Botanic Gardens (Plant DNA C-values)

Table S2. Bisulphite sequencing coverage and alignment statistics.

Species	Replicate	Tissue	Treatment	Sample Name	Total reads			Uniquely mapped reads			Unmapped reads		
					Total	Paired	Single	Total	Paired	Single	Total	Paired	Single
<i>C. rubella</i>	1	root	23°C	ru_M1-C_Rt	287355125	142195725	2963675	168526389	83802688	921013	83699154	40981438	1736278
<i>C. rubella</i>	1	shoot	23°C	ru_M1-C_St	166953496	80888641	5176214	80162038	39851385	459268	65893997	30691421	4511155
<i>C. rubella</i>	1	root	4°C	ru_M1-T_Rt	210026855	103503725	3019405	113765613	56428102	909409	71084604	34646530	1791544
<i>C. rubella</i>	1	shoot	4°C	ru_M1-T_St	191707065	93983770	3739525	80496143	39995461	505221	91833360	44391575	3050210
<i>C. rubella</i>	2	root	23°C	ru_M2-C_Rt	322810562	158564504	5681554	190109847	94622794	864259	93148579	44303133	4542313
<i>C. rubella</i>	2	shoot	23°C	ru_M2-C_St	418027165	203215665	11595835	198794626	98434450	1925726	176428613	83741864	8944885
<i>C. rubella</i>	2	root	4°C	ru_M2-T_Rt	259591261	127577672	4435917	137392463	68137848	1116767	92379664	44722469	2934726
<i>C. rubella</i>	2	shoot	4°C	ru_M2-T_St	240940406	115904369	9131668	108040630	53592665	855300	106234765	49156509	7921747
<i>A. lyrata</i>	1	root	23°C	ly_M1-C_Rt	206907962	101417401	4073160	129114730	64169487	775756	39033148	18062167	2908814
<i>A. lyrata</i>	1	shoot	23°C	ly_M1-C_St	190301038	91519039	7262960	104666042	52041466	583110	32453031	13121636	6209759
<i>A. lyrata</i>	1	root	4°C	ly_M1-T_Rt	278683424	136298837	6085750	160301627	79748047	805533	69034358	32066708	4900942
<i>A. lyrata</i>	1	shoot	4°C	ly_M1-T_St	220405099	108845765	2713569	125436461	62325420	785621	36588538	17616439	1355660
<i>A. lyrata</i>	2	root	23°C	ly_M2-C_Rt	183527687	90341773	2844141	114090382	56719275	651832	34054633	16095143	1864347
<i>A. lyrata</i>	2	shoot	23°C	ly_M2-C_St	171449198	84131825	3185548	85762543	42482376	797791	41366360	19775471	1815418
<i>A. lyrata</i>	2	root	4°C	ly_M2-T_Rt	218781540	107479851	3821838	116194691	57594105	1006481	65021664	31349603	2322458
<i>A. lyrata</i>	2	shoot	4°C	ly_M2-T_St	185385039	91397989	2589061	91580622	45532922	514778	44477092	21399069	1678954
<i>A. thaliana</i>	1	root	23°C	th_M1-C_Rt	185396332	91324347	2747638	125336148	62324030	688088	34649529	16422953	1803623
<i>A. thaliana</i>	1	shoot	23°C	th_M1-C_St	147240664	71744841	3750982	88036731	43768921	498889	32326909	14653923	3019063
<i>A. thaliana</i>	1	root	4°C	th_M1-T_Rt	97901209	48418425	1064359	64952302	32300363	351576	21024943	10211306	602331
<i>A. thaliana</i>	1	shoot	4°C	th_M1-T_St	141802133	70160261	1481611	84695727	42099931	495865	28665902	13964510	736882
<i>A. thaliana</i>	2	root	23°C	th_M2-C_Rt	160912155	77374708	6162739	104993862	52247545	498772	35384989	14943906	5497177
<i>A. thaliana</i>	2	shoot	23°C	th_M2-C_St	156478536	77033551	2411434	82309177	40876938	555301	46662182	22531676	1598830
<i>A. thaliana</i>	2	root	4°C	th_M2-T_Rt	159299891	77943228	3413435	94717513	46972408	772697	45431867	21517160	2397547
<i>A. thaliana</i>	2	shoot	4°C	th_M2-T_St	129559490	63401058	2757374	70398693	34875926	646841	34869443	16524464	1820515

CG		CHG		CHH		C	
Average coverage	Total number	Average coverage	Total number	Average coverage	Total number	Average coverage	Total number
33.8625	3434128	34.2501	3803102	26.9342	19312289	28.8783	26549519
15.7744	3132154	15.4444	3454348	11.972	15839483	13.0379	22425985
26.9226	3400120	26.7205	3766522	20.5784	18842282	22.2972	26008924
15.2546	3213060	15.0614	3551261	11.838	16611322	12.7973	23375643
35.4223	3429172	35.5946	3797200	27.9232	19221553	29.9969	26447925
22.2052	3447134	22.6719	3817694	18.9742	19615891	19.9137	26880719
32.7449	3409521	32.4797	3776484	24.9368	18943868	27.0458	26129873
17.3153	3214718	17.1085	3552618	13.3638	16641064	14.4748	23408400
19.9712	5284802	20.5027	5293127	17.2304	26184961	18.0956	36762890
24.5384	4940032	23.8061	4913937	18.303	22473690	20.0924	32327659
24.5731	5305011	25.009	5312992	20.7736	26411861	21.9256	37029864
21.702	5256794	22.3049	5263098	18.307	25741821	19.3794	36261713
22.9924	5245560	23.3523	5254942	18.956	25752856	20.1773	36253358
17.9552	5080710	18.0419	5080313	14.6466	24001468	15.6436	34162491
24.3672	5250051	24.4961	5258961	19.6583	25819534	21.0392	36328546
18.4384	5136051	18.4444	5141264	14.7632	24582078	15.8476	34859393
35.0745	5476633	34.9213	5999501	27.7819	30370746	29.7599	41846880
25.2109	5332493	25.0427	5837924	19.5087	28401543	21.0935	39571960
18.2912	5447775	18.3395	5969250	15.3966	30033287	16.2008	41450312
25.2141	5379733	25.0776	5893811	19.5244	29070630	21.0943	40344174
28.8727	5470576	28.7867	5992846	23.3503	30290804	24.8541	41754226
20.7603	5361922	20.7857	5872438	16.715	28828250	17.8531	40062610
30.3678	5458510	29.9896	5980288	23.6765	30131728	25.4633	41570526
21.0472	5277195	20.7465	5775220	16.3231	27755732	17.6238	38808147

36-40	41-45	46-50	51-55	56-60	61-65	66-70	71-75	76-80	81-85	86-90	91-95
1.232E-04	1.232E-04	1.232E-04	1.232E-04	1.232E-04	1.232E-04	1.232E-04	1.232E-04	1.232E-04	1.232E-04	1.232E-04	1.232E-04
3.380E-03	2.934E-03	2.983E-03	1.256E-03	1.712E-03	2.890E-03	1.697E-03	1.862E-03	1.903E-03	1.185E-03	1.239E-03	1.882E-03
8.389E-04	5.538E-04	4.347E-04	5.058E-03	9.609E-04	1.092E-03	1.184E-03	1.418E-03	2.020E-03	9.505E-04	3.528E-03	2.827E-03
2.859E-03	8.584E-04	5.639E-04	2.329E-03	2.793E-03	1.445E-03	2.047E-03	9.787E-04	1.523E-03	1.327E-03	3.624E-03	1.156E-03
9.166E-04	9.166E-04	7.361E-04	7.361E-04	7.361E-04	6.466E-04	8.426E-04	4.282E-04	6.325E-04	2.048E-03	2.048E-03	5.637E-03
1.681E-03	7.227E-04	3.459E-04	6.438E-04	1.630E-03	1.110E-03	1.865E-03	2.798E-03	1.897E-03	1.670E-03	1.533E-03	1.334E-03
2.062E-03	1.223E-03	1.307E-03	2.966E-03	8.087E-04	5.919E-03	8.002E-04	7.504E-04	9.168E-04	1.371E-03	3.286E-03	4.384E-03
3.268E-03	1.926E-03	2.300E-03	1.625E-03	2.199E-03	2.233E-03	2.844E-03	2.075E-03	2.167E-03	1.959E-03	2.140E-03	1.433E-03
1.368E-03	1.847E-03	1.559E-03	8.283E-03	8.396E-04	1.222E-03	3.217E-03	8.340E-04	7.919E-04	1.178E-03	1.684E-03	9.628E-04
1.025E-03	1.061E-03	4.775E-03	9.716E-04	6.795E-04	6.867E-04	2.639E-03	2.078E-03	1.696E-03	1.362E-03	2.512E-03	1.197E-03
6.629E-04	1.470E-03	3.683E-04	4.105E-04	1.397E-03	1.572E-03	7.171E-03	1.112E-03	1.984E-03	1.292E-03	4.383E-03	1.412E-03
3.283E-03	1.476E-03	3.479E-03	1.531E-03	2.107E-03	8.879E-04	4.674E-04	6.439E-03	9.772E-04	2.370E-03	4.856E-04	1.826E-03
9.149E-04	9.149E-04	4.312E-04	2.663E-04	4.196E-04	3.729E-04	1.463E-04	3.097E-04	2.671E-03	6.472E-04	2.514E-04	6.660E-03
4.002E-04	4.002E-04	4.002E-04	4.002E-04	4.002E-04	4.002E-04	4.002E-04	4.002E-04	4.002E-04	4.002E-04	4.002E-04	3.964E-04
1.577E-03	4.311E-04	6.535E-03	1.856E-03	8.592E-04	1.221E-03	4.014E-04	2.029E-03	1.491E-03	3.982E-03	2.865E-03	1.656E-03
4.862E-03	1.914E-03	1.733E-03	1.332E-03	3.098E-03	2.929E-03	2.123E-03	3.481E-03	2.794E-03	1.050E-03	1.737E-03	1.944E-03
1.500E-03	1.500E-03	1.655E-03	1.655E-03	1.655E-03	4.144E-04	1.047E-03	1.219E-03	1.930E-03	1.181E-03	1.176E-03	1.781E-03
3.508E-03	2.827E-03	2.124E-03	2.087E-03	1.771E-03	2.749E-03	3.117E-03	2.704E-03	2.647E-03	1.647E-03	2.623E-03	2.408E-03
1.339E-03	1.339E-03	1.502E-03	4.984E-04	9.023E-04	1.593E-03	4.482E-04	1.139E-03	1.679E-03	8.191E-04	6.643E-03	1.710E-03
8.264E-04	2.310E-03	2.102E-03	1.402E-03	2.819E-03	2.074E-03	1.713E-03	1.918E-03	9.756E-04	1.532E-03	2.015E-03	2.709E-03
1.189E-03	1.062E-03	3.223E-04	3.223E-04	4.793E-04	5.574E-04	3.172E-04	1.099E-03	1.304E-03	2.496E-03	1.024E-03	2.588E-03
5.522E-04	5.522E-04	5.522E-04	5.522E-04	5.522E-04	5.522E-04	5.522E-04	7.468E-04	7.468E-04	7.468E-04	7.468E-04	2.451E-04
5.647E-04	5.647E-04	5.647E-04	4.718E-04	4.718E-04	4.718E-04	7.337E-04	7.337E-04	2.393E-04	2.393E-04	4.889E-04	8.917E-04
5.757E-04	5.757E-04	5.757E-04	5.757E-04	5.757E-04	5.757E-04	5.736E-04	5.736E-04	5.736E-04	5.736E-04	8.379E-04	9.124E-04

Coverage bin

96-100	101-105	106-110	111-115	116-120	121-125	126-130	131-135	136-140	141-145	146-150	151-155
1.232E-04	1.232E-04	1.232E-04	1.232E-04	1.232E-04	1.232E-04	1.232E-04	1.232E-04	1.232E-04	1.232E-04	1.232E-04	1.232E-04
1.859E-03	2.369E-03	2.153E-03	1.307E-03	2.940E-03	3.688E-03	7.118E-04	3.620E-03	1.533E-04	0.000E+00	0.000E+00	0.000E+00
1.836E-03	1.777E-03	1.122E-03	1.538E-03	1.523E-03	2.111E-03	1.488E-03	1.064E-03	1.415E-03	3.119E-03	2.937E-03	1.488E-03
2.423E-03	1.214E-03	1.604E-03	1.602E-03	1.394E-03	1.562E-03	2.575E-03	1.185E-03	2.215E-03	1.403E-03	1.586E-04	7.409E-03
3.089E-04	8.00E-05	6.340E-04	8.585E-04	9.926E-04	2.030E-03	1.352E-04	3.540E-03	2.814E-04	2.191E-04	7.507E-04	1.023E-03
2.180E-03	1.293E-03	2.173E-03	2.193E-03	1.569E-03	1.576E-03	2.120E-03	1.890E-03	1.183E-03	3.942E-03	3.382E-04	2.460E-04
1.914E-03	1.211E-03	2.953E-03	8.620E-04	2.068E-03	1.158E-03	1.348E-03	1.265E-03	2.008E-03	2.307E-03	1.606E-03	1.795E-03
2.346E-03	2.276E-03	1.077E-03	2.398E-03	1.740E-03	2.149E-03	1.115E-03	1.635E-03	2.926E-03	2.398E-03	2.758E-04	0.000E+00
1.568E-03	4.809E-03	3.923E-03	1.457E-03	1.322E-03	1.981E-03	1.162E-03	1.460E-03	1.021E-03	1.652E-03	1.420E-03	9.394E-04
2.765E-03	1.262E-03	1.997E-03	1.527E-03	7.782E-04	8.950E-04	2.266E-03	2.131E-03	1.654E-03	1.756E-03	1.788E-03	2.064E-03
1.255E-03	1.421E-03	1.565E-03	1.859E-03	1.752E-03	2.026E-03	1.884E-03	1.595E-03	9.817E-04	1.929E-03	1.648E-03	1.434E-03
3.717E-03	1.130E-03	1.338E-03	2.290E-03	1.040E-03	1.545E-03	1.361E-03	7.689E-04	8.940E-04	9.020E-04	2.275E-03	9.412E-04
4.422E-04	4.244E-04	2.075E-03	3.678E-04	4.659E-03	2.542E-03	1.747E-03	1.025E-03	1.750E-03	3.346E-04	1.227E-03	3.317E-04
3.766E-04	1.580E-04	1.580E-04	1.580E-04	1.239E-04	1.897E-04	4.098E-04	8.698E-04	2.398E-04	3.139E-04	8.160E-04	1.742E-04
2.301E-03	1.882E-03	9.275E-04	1.696E-03	1.939E-03	1.793E-03	1.488E-03	2.630E-03	2.009E-03	1.293E-03	1.232E-03	2.014E-03
3.004E-03	1.519E-03	2.105E-03	1.786E-03	1.658E-03	1.901E-03	2.643E-03	2.293E-03	1.053E-03	2.799E-03	1.582E-03	1.018E-02
1.124E-03	1.103E-03	4.154E-04	5.969E-03	1.046E-03	1.033E-03	1.343E-03	2.263E-03	1.379E-03	3.194E-03	2.051E-03	1.723E-03
2.105E-03	3.136E-03	2.503E-03	2.450E-03	1.607E-03	3.248E-03	3.708E-03	7.706E-04	3.755E-03	2.841E-03	2.539E-04	0.000E+00
8.629E-04	7.467E-04	3.412E-03	2.451E-03	1.025E-03	3.707E-03	3.696E-04	9.536E-04	2.240E-03	8.877E-04	1.064E-03	1.030E-03
1.739E-03	2.089E-03	8.604E-04	2.201E-03	1.360E-03	1.469E-03	1.755E-03	2.257E-03	1.675E-03	3.070E-03	3.541E-03	3.338E-04
4.004E-03	9.047E-04	1.352E-03	6.379E-04	2.138E-03	4.753E-04	2.944E-03	1.577E-03	1.487E-03	1.253E-03	1.156E-03	1.386E-03
6.096E-04	2.034E-04	2.031E-04	1.711E-04	3.559E-04	2.814E-04	4.394E-04	1.869E-04	1.513E-04	1.437E-03	2.022E-04	1.267E-04
1.550E-03	5.147E-04	4.609E-04	2.532E-03	2.073E-03	9.807E-04	2.771E-03	3.071E-03	1.460E-03	3.499E-04	4.863E-04	1.173E-03
3.375E-04	6.560E-04	6.560E-04	4.159E-04	3.079E-04	8.435E-04	1.185E-03	1.339E-04	9.38E-05	2.451E-03	2.451E-03	1.006E-03

156-160	161-165	166-170	171-175	176-180	181-185	186-190	191-195	196-200
1.232E-04	1.085E-04	1.085E-04	4.981E-04	5.18E-05	4.116E-04	1.384E-04	2.816E-04	4.13E-05
0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00
1.341E-03	1.038E-03	8.502E-04	1.564E-03	1.257E-03	1.068E-03	2.293E-03	1.397E-03	0.000E+00
0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00
8.954E-04	1.058E-03	1.348E-03	8.573E-04	8.120E-04	4.437E-04	7.974E-04	5.624E-04	1.437E-03
2.460E-04	2.460E-04	2.460E-04	2.460E-04	2.460E-04	2.460E-04	2.460E-04	2.460E-04	2.460E-04
1.929E-03	1.121E-03	1.581E-03	5.788E-04	1.102E-03	1.328E-03	1.569E-03	1.377E-03	0.000E+00
0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00
1.277E-03	1.707E-03	2.256E-03	1.313E-03	7.149E-04	7.242E-04	6.562E-04	1.412E-03	1.218E-04
2.633E-04	4.282E-03	1.753E-04	2.429E-02	2.429E-02	2.429E-02	2.429E-02	2.429E-02	2.429E-02
1.131E-03	3.865E-03	2.540E-03	4.795E-04	6.566E-03	2.144E-03	1.305E-03	1.305E-03	1.305E-03
5.900E-04	1.294E-03	2.184E-03	8.609E-04	1.648E-03	2.226E-04	0.000E+00	0.000E+00	0.000E+00
1.464E-03	8.336E-04	1.213E-03	1.331E-03	1.052E-03	5.686E-04	9.629E-04	2.101E-03	2.101E-03
2.907E-03	3.21E-05	1.454E-03	1.014E-03	1.403E-04	7.669E-04	2.994E-04	1.888E-04	1.424E-04
1.528E-03	1.635E-03	1.509E-03	1.844E-03	1.557E-03	3.211E-04	2.271E-02	2.271E-02	2.271E-02
0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00
5.947E-04	1.281E-03	8.246E-04	7.089E-04	7.644E-04	1.328E-03	8.888E-04	3.849E-04	6.575E-04
0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00
1.005E-03	9.123E-04	1.912E-03	1.203E-03	9.266E-04	7.442E-04	7.801E-04	1.832E-03	8.05E-05
2.877E-04	0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00
1.122E-03	1.300E-03	1.115E-03	1.727E-03	1.370E-03	1.668E-03	1.227E-03	3.234E-04	1.646E-02
3.382E-03	9.35E-05	1.948E-04	3.694E-04	1.494E-03	3.517E-04	7.046E-04	2.132E-04	8.90E-05
5.565E-04	1.971E-03	1.325E-03	6.082E-04	8.343E-04	4.643E-04	9.535E-04	7.864E-04	3.750E-04
1.140E-04	4.706E-04	1.446E-03	4.523E-04	2.866E-04	3.206E-04	2.488E-04	2.225E-04	2.235E-04

Table S4 A. MR and DMR statistics by sample

Species	Tissue	Treatment	Total MR	Mean MR length	Median MR length	Bases within MR
<i>C. rubella</i>	root	23°C	49282	559.773	275	27586732
<i>C. rubella</i>	shoot	23°C	42465	677.143	333	28754892
<i>C. rubella</i>	root	4°C	49920	538.310	274	26872421
<i>C. rubella</i>	shoot	4°C	48433	569.099	308	27563162
<i>A. lyrata</i>	root	23°C	62430	1044.704	505	65220895
<i>A. lyrata</i>	shoot	23°C	57995	1165.004	545	67564397
<i>A. lyrata</i>	root	4°C	62474	1049.729	506	65580766
<i>A. lyrata</i>	shoot	4°C	54175	1278.907	584	69284797
<i>A. thaliana</i>	root	23°C	28086	824.743	285	23163744
<i>A. thaliana</i>	shoot	23°C	24319	973.973	315	23686061
<i>A. thaliana</i>	root	4°C	28202	819.307	288	23106097
<i>A. thaliana</i>	shoot	4°C	22808	1042.351	342	23773934

Table S4 B. MR and DMR statistics by species

Species	Tissue	Treatment	Total MR	Mean MR length	Median MR length	Bases within MR	Total DMR	Mean DMR length	Median DMR length	Bases within DMR
<i>C. rubella</i>	NA	NA	NA	582.731	296	32137551	8105	237.909	196	1928252
<i>A. lyrata</i>	NA	NA	NA	1128.976	530	73342707	22168	331.097	249	7339750
<i>A. thaliana</i>	NA	NA	NA	906.347	306	25508482	4103	254.822	206	1045535

Table S5. Genome alignment metrics.

Species	Total bases aligned	Total C/G	Conserved context C/G	mC (conserved context)	Conserved mC	Gain mC	Loss mC
<i>C. rubella</i>	58492425	18518838	15138861	314046	117063	92638	40427
<i>A. lyrata</i>	58492425	18518838	15138861	331153	117063	100942	31624
<i>A. thaliana</i>	58492425	18518838	15138861	236977	117063	47863	72721

Table S6. MR presence at genomic features

Species	Category	Exon	5'UTR	Intron	3'UTR	Upstream
<i>C. rubella</i>	all genes	6170	381	2579	715	4033
<i>A. lyrata</i>	all genes	5351	2021	4328	2209	8755
<i>A. thaliana</i>	all genes	5966	300	2027	557	3156
<i>C. rubella</i>	orthologs	2771	128	890	263	1635
<i>A. lyrata</i>	orthologs	348	94	597	196	2564
<i>A. thaliana</i>	orthologs	2775	86	711	221	1128

Table S7. P values of pairwise MR overlap

Comparison	Exon	5'UTR	Intron	3'UTR	Upstream
<i>C. rubella</i> - <i>A. lyrata</i>	4.17E-03	7.81E-01	0.00E+00	1.00E+00	0.00E+00
<i>C. rubella</i> - <i>A. thaliana</i>	0.00E+00	1.74E-04	0.00E+00	3.44E-03	0.00E+00
<i>A. lyrata</i> - <i>A. thaliana</i>	2.56E-02	2.22E-05	3.56E-13	8.06E-04	0.00E+00

Table S8 A. Two and three species MR overlap

	Overlap of MR at orthologs				
	Exon	5'UTR	Intron	3'UTR	Upstream
No MR	9885	14017	12406	13658	10411
MR - 1 species	3004	277	1635	623	2716
MR - 2 species	1370	14	241	27	935
MR - 3 species	50	1	27	1	247

Table S8 B. Permutation values for two and three species MR overlap

	Permutation of overlap of MR at orthologs				
	Exon	5'UTR	Intron	3'UTR	Upstream
No MR	9154	14011	12260	13653	9665
MR - 1 species	4737	308	2048	678	4292
MR - 2 species	707	10	145	24	626
MR - 3 species	28	1	8	2	42

Table S8 C. Data exceeds permutation values for two and three species MR overlap

	Overlap exceeds simulated value				
	Exon	5'UTR	Intron	3'UTR	Upstream
No MR	TRUE	TRUE	TRUE	TRUE	TRUE
MR - 1 species	FALSE	FALSE	FALSE	FALSE	FALSE
MR - 2 species	TRUE	TRUE	TRUE	TRUE	TRUE
MR - 3 species	TRUE	FALSE	TRUE	FALSE	TRUE

Table S9 A. Gene body methylation by context (all genes)

Species	Total gene number	CG methylation	CHG methylation	CHH methylation	CG expressed	CHG expressed	CHH expressed
<i>C. rubella</i>	24129	16718	2501	2544	12274	953	975
<i>A. lyrata</i>	31557	23158	6173	6415	14225	1475	1651
<i>A. thaliana</i>	27387	15733	1880	2348	11963	380	813

Table S9 B. Gene body methylation by context (orthologous genes)

Species	Total ortholog number	CG methylation	CHG methylation	CHH methylation	CG expressed	CHG expressed	CHH expressed
<i>C. rubella</i>	13160	8795	717	744	7311	485	487
<i>A. lyrata</i>	13160	9635	954	1166	7939	686	841
<i>A. thaliana</i>	13160	7617	320	581	6506	171	413

Table S10. Three-way genome alignment site classes by context

Species	Context	Aligned	All mC	3-way mC	2-way mC	Gain mC	Loss mC
<i>C. rubella</i>	CG	1634764	288600	116786	103181	68633	39683
<i>C. rubella</i>	CHG	2247539	8844	132	462	8250	336
<i>C. rubella</i>	CHH	11256558	16602	145	702	15755	408
<i>A. lyrata</i>	CG	1634764	305599	116786	111862	76951	31002
<i>A. lyrata</i>	CHG	2247539	8630	132	538	7960	260
<i>A. lyrata</i>	CHH	11256558	16924	145	748	16031	362
<i>A. thaliana</i>	CG	1634764	222366	116786	70685	34895	72179
<i>A. thaliana</i>	CHG	2247539	5466	132	596	4738	202
<i>A. thaliana</i>	CHH	11256558	9145	145	770	8230	340

Table S11. DMP statistics by comparison

Species	Total	Tissue (23C)	Tissue (4C)	Treatment (root)	Treatment (shoot)	Tissue (both)	Treatment (both)	Other
<i>C. rubella</i>	464191	180419	205857	18089	29589	23336	16	6885
<i>A. lyrata</i>	1442130	686091	570122	39827	47609	93028	7	5446
<i>A. thaliana</i>	125423	55853	48911	5177	4849	10390	2	241

Table S12. DMR statistics by comparison

Species	Total	Tissue	Treatment
<i>C. rubella</i>	8105	5700	288
<i>A. lyrata</i>	22168	18613	395
<i>A. thaliana</i>	4103	3210	74

Table S13. DMR presence at genomic features

Species	Category	Exon	5'UTR	Intron	3'UTR	Upstream	
<i>C. rubella</i>	all genes		405	17	298	54	671
<i>A. lyrata</i>	all genes		1341	243	1188	273	1836
<i>A. thaliana</i>	all genes		281	24	159	34	393
<i>C. rubella</i>	orthologs		90	4	86	22	244
<i>A. lyrata</i>	orthologs		94	7	176	35	437
<i>A. thaliana</i>	orthologs		76	4	40	12	125

Table S14. P values of pairwise DMR overlap

Comparison	Exon	5'UTR	Intron	3'UTR	Upstream
<i>C. rubella</i> - <i>A. lyrata</i>	3.21E-01	1.00E+00	1.57E-04	1.00E+00	1.10E-07
<i>C. rubella</i> - <i>A. thaliana</i>	1.00E+00	1.00E+00	2.65E-02	1.00E+00	8.38E-02
<i>A. lyrata</i> - <i>A. thaliana</i>	2.34E-02	1.00E+00	1.29E-04	4.35E-01	2.32E-02

Table S15 A. Two and three species DMR overlap

	Overlap of DMR at orthologs				
	Exon	5'UTR	Intron	3'UTR	Upstream
No DMR	14056	14294	14024	14241	13547
DMR - 1 species	246	15	268	67	719
DMR - 2 species	7	0	17	1	42
DMR - 3 species	0	0	0	0	1

Table S15 B. Permutation values for two and three species DMR overlap

	Simulation of overlap of DMR at orthologs				
	Exon	5'UTR	Intron	3'UTR	Upstream
No DMR	14057	14295	14016	14243	13533
DMR - 1 species	260	15	302	69	802
DMR - 2 species	8	1	9	3	30
DMR - 3 species	1	0	1	0	2

Table S15C. Data exceeds permutation values for two and three species DMR overlap

	Overlap exceeds simulated value				
	Exon	5'UTR	Intron	3'UTR	Upstream
No DMR	FALSE	FALSE	TRUE	FALSE	TRUE
DMR - 1 species	FALSE	FALSE	FALSE	FALSE	FALSE
DMR - 2 species	FALSE	FALSE	TRUE	FALSE	TRUE
DMR - 3 species	FALSE	FALSE	FALSE	FALSE	FALSE

Table S16. DMPs at genomic features

Species	Category	Exon	5'UTR	Intron	3'UTR	Upstream
<i>C. rubella</i>	all genes	10584	230	3926	827	2977
<i>A. lyrata</i>	all genes	17549	672	7835	1096	7625
<i>A. thaliana</i>	all genes	7412	93	1875	335	1695
<i>C. rubella</i>	orthologs	5402	45	1841	370	1153
<i>A. lyrata</i>	orthologs	7529	43	2778	269	2162
<i>A. thaliana</i>	orthologs	3758	23	793	145	554

Table S17. DMP correlation with gene expression by feature

Species	Comparison	Exon	Intron	Upstream	5'UTR	3'UTR
<i>C. rubella</i>	Tissue	-0.0817	0.0061	0.1084	-0.0745	0.0070
<i>A. lyrata</i>	Tissue	-0.0646	-0.0318	0.0913	-0.1468	-0.0214
<i>A. thaliana</i>	Tissue	-0.1008	0.0014	-0.0188	-0.1748	-0.0487
<i>C. rubella</i>	Treatment	0.0194	0.0203	0.0588	0.1313	0.0642
<i>A. lyrata</i>	Treatment	-0.0604	0.0471	-0.0099	0.0000	-0.0044
<i>A. thaliana</i>	Treatment	0.0362	0.0390	-0.0333	NA	0.0111

Table S18. DMR correlation with gene expression by context

Species	Context	CDS	intron	upstream
<i>C. rubella</i>	CG	-0.1782	-0.1948	-0.1177
<i>A. lyrata</i>	CG	-0.2416	-0.0684	-0.0859
<i>A. thaliana</i>	CG	-0.1518	-0.2447	-0.1747
<i>C. rubella</i>	CHG	-0.2744	0.0326	0.1460
<i>A. lyrata</i>	CHG	-0.0292	0.0770	0.0290
<i>A. thaliana</i>	CHG	-0.4585	-0.1367	0.0987
<i>C. rubella</i>	CHH	-0.2673	0.0343	0.2184
<i>A. lyrata</i>	CHH	0.1244	0.0266	0.0081
<i>A. thaliana</i>	CHH	-0.3753	-0.0296	0.2489

Table S19. Number of DMPs between replicates

Species	Tissue	Treatment	Context	Number of DMPs
<i>C. rubella</i>	root	23°C	CG	48
<i>C. rubella</i>	shoot	23°C	CG	260
<i>C. rubella</i>	root	4°C	CG	27
<i>C. rubella</i>	shoot	4°C	CG	28
<i>C. rubella</i>	root	23°C	CHG	26
<i>C. rubella</i>	shoot	23°C	CHG	372
<i>C. rubella</i>	root	4°C	CHG	33
<i>C. rubella</i>	shoot	4°C	CHG	56
<i>C. rubella</i>	root	23°C	CHH	196
<i>C. rubella</i>	shoot	23°C	CHH	9012
<i>C. rubella</i>	root	4°C	CHH	269
<i>C. rubella</i>	shoot	4°C	CHH	690
<i>A. lyrata</i>	root	23°C	CG	5845
<i>A. lyrata</i>	shoot	23°C	CG	11209
<i>A. lyrata</i>	root	4°C	CG	212703
<i>A. lyrata</i>	shoot	4°C	CG	6314
<i>A. lyrata</i>	root	23°C	CHG	911
<i>A. lyrata</i>	shoot	23°C	CHG	1260
<i>A. lyrata</i>	root	4°C	CHG	21232
<i>A. lyrata</i>	shoot	4°C	CHG	688
<i>A. lyrata</i>	root	23°C	CHH	350
<i>A. lyrata</i>	shoot	23°C	CHH	799
<i>A. lyrata</i>	root	4°C	CHH	8393
<i>A. lyrata</i>	shoot	4°C	CHH	287
<i>A. thaliana</i>	root	23°C	CG	2
<i>A. thaliana</i>	shoot	23°C	CG	1
<i>A. thaliana</i>	root	4°C	CG	3
<i>A. thaliana</i>	shoot	4°C	CG	5
<i>A. thaliana</i>	root	23°C	CHG	0
<i>A. thaliana</i>	shoot	23°C	CHG	0
<i>A. thaliana</i>	root	4°C	CHG	0
<i>A. thaliana</i>	shoot	4°C	CHG	1
<i>A. thaliana</i>	root	23°C	CHH	0
<i>A. thaliana</i>	shoot	23°C	CHH	0
<i>A. thaliana</i>	root	4°C	CHH	0
<i>A. thaliana</i>	shoot	4°C	CHH	0

Table S20. RNA-seq coverage and alignment statistics

Species	Replicate	Tissue	Treatment	Sample Name	Read counts	Mapped reads	Uniquely mapped reads	Pass quality threshold (30)
<i>C. rubella</i>	1	root	23°C	ruR1Crt	5318062	3112856	3084205	2606363
<i>C. rubella</i>	1	shoot	23°C	ruR1Cst	7336247	5101865	5017804	4490465
<i>C. rubella</i>	1	root	4°C	ruR1Trt	5494219	3233928	3206313	2710305
<i>C. rubella</i>	1	shoot	4°C	ruR1Tst	13420000	8663361	8514518	7238646
<i>C. rubella</i>	2	root	23°C	ruR2Crt	7905410	4818010	4771033	4042079
<i>C. rubella</i>	2	shoot	23°C	ruR2Cst	13229689	9242844	9111119	7735911
<i>C. rubella</i>	2	root	4°C	ruR2Trt	6418070	3616100	3581960	3072624
<i>C. rubella</i>	2	shoot	4°C	ruR2Tst	6280040	4131658	4059935	3464742
<i>C. rubella</i>	3	root	23°C	ruR3Crt	9112065	5523276	5472680	4938805
<i>C. rubella</i>	3	shoot	23°C	ruR3Cst	17805382	12127969	11926547	10677571
<i>C. rubella</i>	3	root	4°C	ruR3Trt	7287573	4209767	4169996	3512520
<i>C. rubella</i>	3	shoot	4°C	ruR3Tst	6546173	4270986	4194464	3510736
<i>A. lyrata</i>	1	root	23°C	lyR1Crt	10560799	6224563	6053207	5343928
<i>A. lyrata</i>	1	shoot	23°C	lyR1Cst	6919860	4420348	4263273	3769132
<i>A. lyrata</i>	1	root	4°C	lyR1Trt	6196153	3148461	3071869	2540094
<i>A. lyrata</i>	1	shoot	4°C	lyR1Tst	10898970	6794899	6595828	5492818
<i>A. lyrata</i>	2	root	23°C	lyR2Crt	6191182	3552837	3455620	2901415
<i>A. lyrata</i>	2	shoot	23°C	lyR2Cst	7931392	2649858	2510751	2161840
<i>A. lyrata</i>	2	root	4°C	lyR2Trt	5430782	3123750	3047206	2581982
<i>A. lyrata</i>	2	shoot	4°C	lyR2Tst	9935412	6434788	6246099	5265106
<i>A. lyrata</i>	3	root	23°C	lyR3Crt	14979376	9230912	8971543	7966199
<i>A. lyrata</i>	3	shoot	23°C	lyR3Cst	5699244	3538052	3415047	2823457
<i>A. lyrata</i>	3	root	4°C	lyR3Trt	9407678	5572531	5433835	4540386
<i>A. lyrata</i>	3	shoot	4°C	lyR3Tst	8641565	5648010	5477328	4866095
<i>A. thaliana</i>	1	root	23°C	thR1Crt	10820919	6689126	6614770	5969343
<i>A. thaliana</i>	1	shoot	23°C	thR1Cst	13626140	9084147	8962521	7557086
<i>A. thaliana</i>	1	root	4°C	thR1Trt	8356957	5059712	5019260	4498031
<i>A. thaliana</i>	1	shoot	4°C	thR1Tst	10251950	7207721	7121574	6128815
<i>A. thaliana</i>	2	root	23°C	thR2Crt	12679579	8257305	8162827	7050999
<i>A. thaliana</i>	2	shoot	23°C	thR2Cst	8305847	5867401	5786312	5203513
<i>A. thaliana</i>	2	root	4°C	thR2Trt	5866246	1996604	1980277	1697763
<i>A. thaliana</i>	2	shoot	4°C	thR2Tst	16614616	11733298	11593720	9824714
<i>A. thaliana</i>	3	root	23°C	thR3Crt	13675597	9146105	9042668	7777927
<i>A. thaliana</i>	3	shoot	23°C	thR3Cst	16323055	11320801	11165099	9464125
<i>A. thaliana</i>	3	root	4°C	thR3Trt	7245439	4796007	4759049	4302689
<i>A. thaliana</i>	3	shoot	4°C	thR3Tst	13206825	9108865	8996200	7712869

```
#####  
Command Lines for processing BS reads  
#####
```

```
### Filter reads with SHORE v0.8:  
shore import -a genomic -b <folder_with_original_qseq_readfiles> -B  
1000000 --rplot -n 10% -g -c -k 40 --discard-trim-failures -h 2 -w  
2ndread -o <output_folder> -r <barcode_file> -l <lane_nr>
```

```
### Mapping with bismark v0.7.3:  
bismark --seedlen 40 --seedmms 2 --maqerr 100 --chunkmbs 1024 --minins  
100 --maxins 1000 --ambiguous <reference_genome_folder> -1 reads_1.fq  
-2 reads_2.fq  
bismark --seedlen 40 --seedmms 2 --maqerr 100 --chunkmbs 1024 --  
ambiguous <reference_genome_folder> reads_single.fq
```

```
### Paired-end correction:  
shore correct4pe -e 1 -l <sample_folder> -x 300
```

```
### Retrieving read counts:  
shore methyl -n <sampleID> -f <reference_file> -o <output_dir> -i  
<samples_SAM_file> -b -a scoring_matrix_meth.txt -g 0 -m 0
```

```
#####
```

```
scoring_matrix_meth.txt:  
support_core      3      4      5  
support 3         4      5  
support_inner_core 0      0      0  
quality_max       20     30     35  
quality_diff      20     10     5  
startpos_core    3      4      5  
cn_noise          0.3    0.15   0.05  
quality_max_noncall 99     35     30  
avg_hits          2      1.5    1.1  
percent_mm        0.10   0.07   0.05  
exp_obs 8.0       5.0    3.0  
depth_obs         10.0   7.0    5.0  
gc_cont 5         10     15  
seq_complexity    0      1      3  
pc                5      8      10  
fwd_rev_ratio     30     10     5  
left_right_ratio  30     10     5
```



```
#####  
Command Lines for processing RNA-seq reads  
#####
```

```
### Filter reads with SHORE v0.8:
```

```
shore import -a mRNA -b <folder_with_original_qseq_readfiles> -l  
<lane_nr> -o <output_folder> -n 10 -c -g -k 75 -r <barcode_file> -h 1 -  
w 2ndRead --discard-trim-failures --rplot
```

```
### Mapping with bwa v0.6.1:
```

```
bwa aln -n 1 -t 4 <alignment_index> <sample_reads.fq> > <output.sai>
```

```
### Decompress bwa .sai file:
```

```
bwa samse <alignment_index> <sample.sai> <sample_reads.fq> >  
<output.sam>
```

4. “Improving the annotation of *Arabidopsis lyrata* using RNA-seq data”

Rawat V, Abdelsamad A, Pietzenuk B, Seymour DK, Koenig D, Weigel D, Pecinka A, Schneeberger K (2015).

PLoS One 10:e0137391. doi: 10.1371/journal.pone.0137391.

Abstract

Gene model annotations are important community resources that ensure comparability and reproducibility of analyses and are typically the first step for functional annotation of genomic regions. Without up-to-date genome annotations, genome sequences cannot be used to maximum advantage. It is therefore essential to regularly update gene annotations by integrating the latest information to guarantee that reference annotations can remain a common basis for various types of analyses. Here, we report an improvement of the *Arabidopsis lyrata* gene annotation using extensive RNA-seq data. This new annotation consists of 31,132 protein coding gene models in addition to 2,089 genes with high similarity to transposable elements. Overall, ~87% of the gene models are corroborated by evidence of expression and 2,235 of these models feature multiple transcripts. Our updated gene annotation corrects hundreds of incorrectly split or merged gene models in the original annotation, and as a result the identification of alternative splicing events and differential isoform usage are vastly improved.

Contributions

Conceived and designed the experiments: KS AP. Performed the experiments: AP AA. Analyzed the data: VR. Contributed reagents/materials/analysis tools: AA BP DKS DK DW. Wrote the paper: VR KS AA AP BP DKS DK DW.

License

“PLOS applies the Creative Commons Attribution (CC BY) license to works we publish. This license was developed to facilitate open access – namely, free immediate access to, and unrestricted reuse of, original works of all types. Under this license, authors agree to make articles legally available for reuse, without permission or fees, for virtually any purpose. Anyone may copy, distribute or reuse these articles, as long as the author and original source are properly cited.”

<http://journals.plos.org/plosone/s/content-license>

RESEARCH ARTICLE

Improving the Annotation of *Arabidopsis lyrata* Using RNA-Seq Data

Vimal Rawat¹, Ahmed Abdelsamad^{2,3}, Björn Pietzenek², Danelle K. Seymour⁴, Daniel Koenig⁴, Detlef Weigel⁴, Ales Pecinka^{2*}, Korbinian Schneeberger^{1*}

1 Department of Plant Developmental Biology, Max Planck Institute for Plant Breeding Research, Carl-von-Linné-Weg 10, 50829, Cologne, Germany, **2** Department of Plant Breeding and Genetics, Max Planck Institute for Plant Breeding Research, Carl-von-Linné-Weg 10, 50829, Cologne, Germany, **3** Department of Genetics, Cairo University, 12613, Giza, Egypt, **4** Department for Molecular Biology, Max Planck Institute for Developmental Biology, Spemannstrasse 35–39, 72076, Tübingen, Germany

* schneeberger@mpipz.mpg.de (KS); pecinka@mpipz.mpg.de (AP)



OPEN ACCESS

Citation: Rawat V, Abdelsamad A, Pietzenek B, Seymour DK, Koenig D, Weigel D, et al. (2015) Improving the Annotation of *Arabidopsis lyrata* Using RNA-Seq Data. PLoS ONE 10(9): e0137391. doi:10.1371/journal.pone.0137391

Editor: Nicholas James Provart, University of Toronto, CANADA

Received: June 12, 2015

Accepted: August 17, 2015

Published: September 18, 2015

Copyright: © 2015 Rawat et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: Short read data presented in this paper are available through NCBI Short Read Archive under accession number GSE69222. Short reads from the cold treatment are deposited at the European Nucleotide Archive under accession number PRJEB6701.

Funding: This work was funded by German Research Foundation (dfg.de) grant AP1829-2 to AP. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Gene model annotations are important community resources that ensure comparability and reproducibility of analyses and are typically the first step for functional annotation of genomic regions. Without up-to-date genome annotations, genome sequences cannot be used to maximum advantage. It is therefore essential to regularly update gene annotations by integrating the latest information to guarantee that reference annotations can remain a common basis for various types of analyses. Here, we report an improvement of the *Arabidopsis lyrata* gene annotation using extensive RNA-seq data. This new annotation consists of 31,132 protein coding gene models in addition to 2,089 genes with high similarity to transposable elements. Overall, ~87% of the gene models are corroborated by evidence of expression and 2,235 of these models feature multiple transcripts. Our updated gene annotation corrects hundreds of incorrectly split or merged gene models in the original annotation, and as a result the identification of alternative splicing events and differential isoform usage are vastly improved.

Introduction

Arabidopsis lyrata is a predominantly self-incompatible, perennial plant species that diverged from a common ancestor with *A. thaliana* approximately 10 million years ago [1]. Despite its evolutionary closeness, its genome size is estimated to be between 230 to 245 Mb, or one and a half times as large as the *A. thaliana* genome [2,3]. Except for *A. thaliana*, *A. lyrata* is the only species within the family of Brassicaceae with a reference assembly exclusively based on high quality dideoxy sequencing. This 207 Mb *A. lyrata* reference assembly attributed the difference in genome size between the two species to the accumulation of many small deletions in the *A. thaliana* genome primarily in non-coding regions and transposable elements (TEs) [1]. In addition, *A. lyrata* has experienced recent genome expansion due to activity of TEs, in particular Copia long terminal repeat (LTR) retrotransposons [1,4,5], which is the basis for species-specific patterns in DNA methylation [6].

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

As *A. lyrata* is the closest fully assembled relative of *A. thaliana*, it serves as an important out-group for evolutionary studies within *A. thaliana* [7–9]. Moreover, recent advances in sequencing technology have enabled the assembly of an increasing number of Brassicaceae genomes and their close relatives [4,5,10–19], which, together, are leveraged for comparative genomics in this family. Intra- as well as inter-species comparisons, however, heavily rely on the gene annotations of each species involved and high quality annotations even in the non-model species become essential.

Methods for gene model annotation profited considerably from the invention of high-throughput RNA sequencing (RNA-seq) [20,21]. Identification of genuine transcription start and termination sites as well as intron/exon borders is a non-trivial task when using only reference sequences and homology data. Now, information on spliced alignments from RNA-seq data can improve the identification of gene models [22,23] and also enable the annotation of variant isoforms [24]. In particular, the gene annotations of model species have been updated regularly despite only minor changes to the reference genome sequence [25].

The current gene annotation of the *A. lyrata* includes 32,670 genes and was generated using a combination of *ab initio* gene prediction, homology to known proteins, as well as gene sequences and expression data from related species [1]. Even though the gene models were analyzed for their expression support using RNA-seq data, gene prediction methods integrating RNA-seq alignment information had not been developed at the time the assembly had been generated. In a recent study, Haudry and colleagues supplemented the original annotation with additional putatively transcribed regions in order to study the conservation of non-coding sequences among related Brassicaceae species [14]. They integrated the results of additional *ab initio* gene predictions, RNA-seq data alignments and homology searches against the genes of *A. thaliana* in order to mask potentially un-annotated coding sequences and regions that recently lost coding potential due to mutations.

Building upon the major efforts of the first annotation of *A. lyrata* genome (version-1 from hereon) we have updated the gene models using diverse RNA-seq samples. Our annotation (version-2 from hereon) has changed the coordinates of 29,141 of the original 32,670 gene models, removed 1,286 and added 1,295 new models. This update corrected hundreds of gene models, which were wrongly merged or split in version-1, and also separated transposable element genes from other protein coding genes. Finally, we have analyzed the transcriptional response of *A. lyrata* to heat stress to show the improved utility of version-2 for the identification of differential isoform usage and pre-mRNA splicing.

Results and Discussion

Improving the *A. lyrata* gene annotation using transcriptional data

We sequenced the transcriptome of various *A. lyrata* aerial tissues, including whole rosettes, dissected shoot apices, complete inflorescences, as well as vegetative rosettes exposed to cold and heat stress (see [Materials and Methods](#)). In total, we generated over 290 million single-end, strand unspecific short reads using Illumina sequencing technology after poly-A purification ([Table 1](#)). Short reads were aligned to the *A. lyrata* reference assembly [1] using Bowtie v2.1.0 [26] and the splice junction mapper TopHat v2.0.9 [27] (see [Materials and Methods](#)). We could align 89% of all reads, out of which 85% aligned uniquely and were used for further analysis. The proportion of unaligned reads was comparable to the proportion of unaligned reads in similar experiments with *A. thaliana*, which presumably has one of the most complete reference genome sequences. Over 10% of the aligned reads matched to putative intergenic regions indicating that some gene models may have been missed in the original version of the *A. lyrata* gene annotation. Visual inspection of these intergenic alignments revealed the expected

Table 1. Short read statistics (read numbers in millions).

Tissue	Sample	Read length (bp)	Raw reads	Reads aligned	Reads aligned uniquely (full-length alignments)	Reads aligned uniquely (spliced alignments)
Rosette (WT)	Rep 1	96	16.0	11.9	8.3	3.0
Rosette (WT)	Rep 2	96	12.0	9.1	6.0	2.7
Rosette (Heat stressed)	Rep 1	96	17.6	13.1	9.7	3.4
Rosette (Heat stressed)	Rep 2	96	5.9	4.6	3.1	1.2
Rosette (Recovered)	Rep 1	96	15.6	12.6	5.6	4.1
Rosette (Recovered)	Rep 2	96	5.8	4.5	2.8	1.3
Shoot apical meristem (WT)	Rep 1	101	14.5	12.7	7.5	3.6
Rosette (WT)	Rep 1	101	19.6	17.5	9.3	5.6
Rosette (WT)	Rep 2	101	18.1	16.2	9.3	4.8
Inflorescence (WT)	Rep 1	75	32.0	30.2	12.9	9.3
Inflorescence (WT)	Rep 2	75	32.0	29.5	12.6	9.0
Rosette (WT and cold stressed)	Rep 1	75–100	102.7	96.7	59.7	24.6
			291.8	258.6	146.8	72.6

doi:10.1371/journal.pone.0137391.t001

patterns for spliced transcripts indicating instances of unidentified gene models and cases where transcription exceeded known gene boundaries (see Fig A in [S1 Dataset](#)).

New gene models were predicted from short read alignment data using Cufflinks 2.1.1 [22] independently for each tissue. In total, Cufflinks predicted 31,194 distinct gene models across all samples. An additional RNA-seq alignment-guided gene prediction using Augustus v.3.0.1 [28] identified 40,728 gene models, including 27,830 genes, which were supported by at least five RNA-seq reads. Moreover, 30,483 and 30,837 of Augustus predicted gene models overlapped with version-1 and Cufflinks predictions, respectively (see [Materials and Methods](#) and Fig B in [S1 Dataset](#)).

We combined 31,793 Augustus gene models with evidence of transcription or that overlapped with version-1 gene models to update the *A. lyrata* gene annotation (Fig 1A). To ensure that we were not excluding any true gene models in version-1, we included 1,430 version-1 gene models that were not overlapping with any of the new gene models, but showed either evidence of expression or featured an ortholog in at least one of the Brassicaceae species *A. thaliana* [29], *Capsella rubella* [4], *Brassica rapa* [10], *Schrenkiella parvula* [11] and *Arabis alpina* [5]. This increased the number of gene models to 33,223 (see [Materials and Methods](#)). To identify and to correct cases where incorrect gene models may have been introduced into the version-2 annotation, we utilized the very close phylogenetic relationship between *A. lyrata*, *A. thaliana* and *C. rubella*. We compared all gene models that were considerably different between version-1 and version-2 to *A. thaliana* and *C. rubella* orthologs (see [Materials and Methods](#)). If the length of the version-1 open reading frame was closer to that of the orthologs, we retained the version-1 gene model. This resulted in 548 version-2 gene models being replaced with 688 of the original version-1 gene models (Fig 1B). After additional removal of redundant gene models we obtained a final set of 33,221 non-redundant gene models.

Based on a recent annotation of *A. lyrata* TEs [14] and sequence similarity to TE genes of *A. thaliana* [25], we annotated 2,089 of the protein coding gene models as TE protein coding genes (see [Materials and Methods](#)). Without these, version-2 comprised of 31,132 gene models, which is ~13% more than in *A. thaliana* [25]. Although tRNA genes were described in the original analysis of the *A. lyrata* genome [1], version-1 lacks information regarding these loci. By

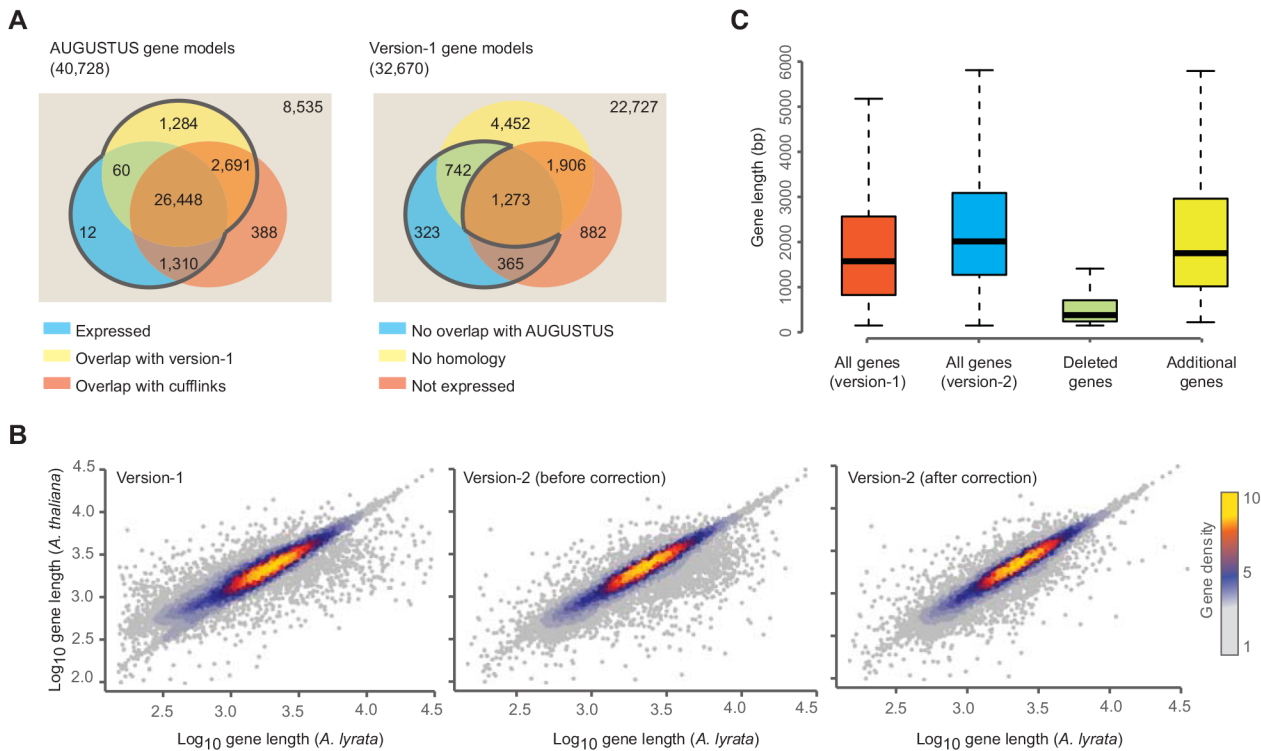


Fig 1. Updating the gene model annotation of *A. lyrata*. (A) Left, version-2 gene models predicted by Augustus [28]. Number of gene models overlapping with version-1 (yellow), genes predicted with Cufflinks (red), and genes with expression evidence (blue). Right, gene models of the version-1 annotation. Number of models without overlap to version-2 models (yellow), without orthologs in five other Brassicaceae (red), and without significant expression evidence (blue). (B) Correlation of the lengths of *A. lyrata* gene models with the length of their orthologous gene models in *A. thaliana*. Left, *A. lyrata* version-1 gene models. Correlations using version-1 gene models (left), version-2 gene models before (middle) and after (right) the homology-based correction of gene models. (C) Length distribution of gene models including genes that were removed or newly added in the version-2.

doi:10.1371/journal.pone.0137391.g001

rerunning tRNAScan [30], we identified 660 tRNA genes coding for all 20 amino acids. For completeness, we also incorporated 170 recently published miRNA genes into the new annotation file [31].

Altogether, we updated the coordinates of 29,141 of the original gene models, removed 1,286 entire (mostly short) gene models, and added 1,295 new models (Fig 1C). Only 2,243 remained unaltered (including 688 version-1 gene models re-introduced due to their superior similarity to orthologs). The new annotation accounted for 31,132 non-TE related gene models including 27,084 multi-exonic genes of which 2,236 featured at least one alternative isoform (Table 2).

Table 2. Comparison of version-1 and version-2 annotations.

	# version-1	# version-2
Gene models	32,670	33,221
Predicted transcripts	32,670	35,805
Protein coding genes	32,670	31,132
TE coding genes	-	2,089
miRNA genes	-	170
tRNA genes	-	660
Featuring ortholog (in at least one Brassicaceae)	23,996	24,146

doi:10.1371/journal.pone.0137391.t002

Validating differences in gene model structure

Even after the above-mentioned homology-based gene length adjustments, we found cases where the corresponding gene models from the two annotations varied drastically in length. This included instances where multiple version-1 gene models were fused to form a single gene model in version-2 or vice versa (Fig 2). In total, 161 version-1 genes were split (accounting for 530 genes in version-2) and 1,729 version-1 gene models were merged (accounting for 775 gene models in version-2). We randomly selected 14 version-1 gene models that had been split into multiple gene models in version-2, and another 14 gene models that had been merged in version-2, for PCR validation (see Fig C and D in S1 Dataset). One split case did not yield gDNA bands indicating a technical problem in primer design. For three merge cases we obtained cDNA bands of the expected size, but were not able to amplify genomic DNA for primer validation. This was most likely due to large gDNA amplicon size (2.4–5 kbp) and

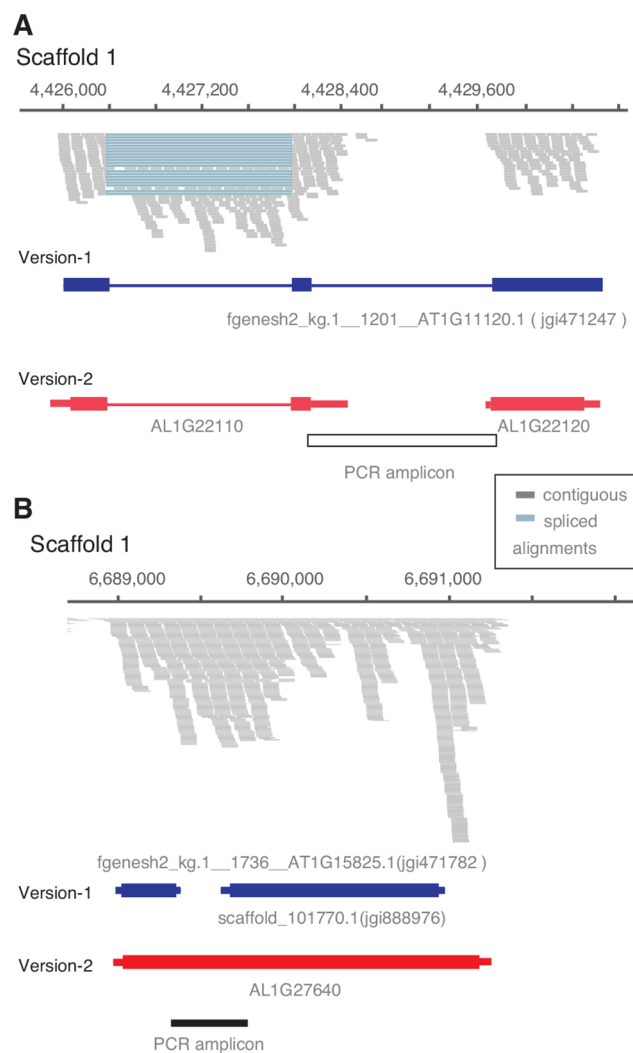


Fig 2. Examples of version-1 gene models split and merged in *A. lyrata* gene annotation version-2. (A) Example of a gene model that was split into two gene models in version-2. Reverse transcription-PCR could not confirm the connection of both. **(B)** Example of version-1 gene models that were merged during the annotation update. Reverse transcription-PCR confirmed presence of a transcript bridging the two version-1 genes.

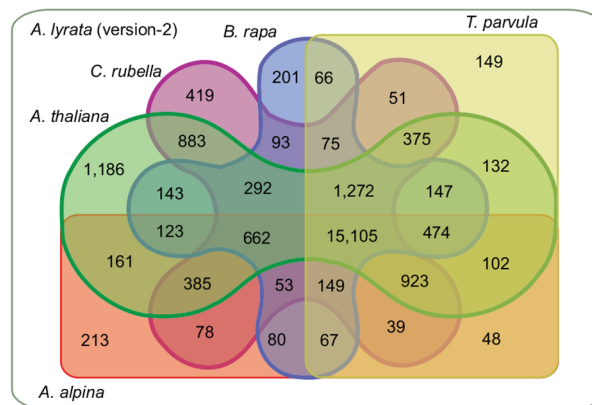
doi:10.1371/journal.pone.0137391.g002

rendered the results of these cases inconclusive. For all 24 remaining cases, PCR results fully confirmed the annotation of the new gene models.

A. lyrata version-2 annotation in contrast to other Brassicaceae annotations

For both *A. lyrata* annotations we predicted orthologous relationships between *A. lyrata* and five other Brassicaceae species (see [Materials and Methods](#)). Using version-2 gene models, 77.5% of genes had an ortholog in at least one species (24,146 out of 31,132) ([Fig 3A](#)), compared to 73% for version-1 (23,996 out of 32,670) (see Fig E in [S1 Dataset](#)). The number of genes with orthologs in all five Brassicaceae was also slightly higher for version-2 with 15,105 genes versus 14,850 genes with version-1.

A



B

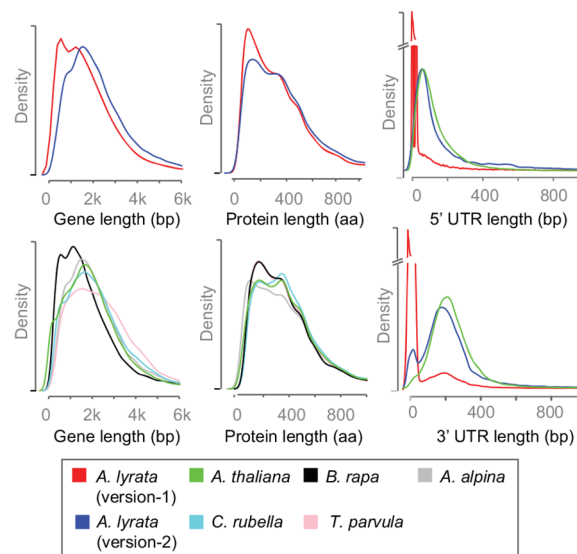


Fig 3. Comparing the *A. lyrata* gene annotation version-2 with the annotations of five other Brassicaceae. (A) Orthologous gene models shared between *A. lyrata* (version-2), *A. thaliana* [29], *A. alpina* [5], *B. rapa* [10], *C. rubella* [4] and *S. parvula* [11]. (B) Gene, Protein and UTR length distributions of above-mentioned species including the new and old *A. lyrata* annotations. UTR distribution is only shown for *A. lyrata* and *A. thaliana* because of poor UTR annotation in some of the other species.

doi:10.1371/journal.pone.0137391.g003

The removal of many short gene models in version-2 changed the distribution of gene model lengths (Figs 1C and 3B). Version-1 has an excess of gene models shorter than 1 kb with a second mode around 1.5 kb, which describes a bimodal distribution that was only reflected by gene length distribution of *B. rapa*. In contrast version-2 had only a single mode around 1.7 kb, similar to the four other species. The length distribution of predicted protein sequences in version-1 had also been distinct from the other Brassicaceae species, and this discrepancy largely disappeared with version-2. A third factor that contributed to the length differences between the genes of version-1 and version-2 were differences in UTR annotations (Fig 3B). In version-1 33% of the genes were annotated without UTR information, however, in version-2 only 5% remained without 3' and 5' UTR annotation. The absolute and relative contributions of individual features are shown in Fig F in S1 Dataset. Though, absolute increase in genomic space for all gene features was observed but CDS and UTRs are benefited the most. We also observed little decrease in intronic genome space, which can be explained by introduction of splice variants previously missing from version-1 annotation.

Whether the bimodal distribution in *B. rapa* reflects similar ambiguity in gene annotations, or mirrors particular characteristics of *B. rapa*, including its ancient genome triplication and subsequent fractionation, is not known.

New annotation enabled improved identification of alternative splicing events

The availability of multiple isoforms from individual gene models in version-2 enables quantitative expression comparisons between annotated isoforms. We analyzed RNA-seq data from *A. lyrata* rosette tissues from untreated (WT), heat stressed (HS), and recovered (REC) samples in duplicate (see Materials and Methods). We first analyzed the data for differential gene expression using Cuffdiff v.2.0.2 [32]. WT and REC differed from HS at 3,114 and 2,962 genes, whereas only 106 genes differed between WT and REC. This indicates, as expected, a strong effect of heat stress on gene expression (see Materials and Methods). Cuffdiff was also used to estimate differential expression between isoforms. We identified differential isoform expression at 283, 15 and 119 genes when comparing WT with HS, WT with REC, and HS with REC, respectively. In contrast, as version-1 does not include different isoforms, which are prerequisite for isoform expression analysis as implemented in Cuffdiff, it was not possible run this analysis using version-1.

We investigated differential splicing using a second tool, MATS v3.0.8 [33], which does not rely on prior isoform annotations and only identifies differences in individual splicing events, but not between entire transcripts. With version-2, MATS identified 177, 0 and 130 differential splicing events distributed over 187 distinct gene models in the three comparisons (Fig 4; see Materials and Methods). MATS reported only 99, 1 and 67 events affecting 103 gene models using version-1. The overlap of different splicing events was very high (95 out of 103 (version-1) and 187 (version-2) gene models). Thus, almost all gene models with differential splicing events predicted based on version-1 were also predicted using version-2, however, the results based on version-2 revealed many more gene models. This was partially due to newly added genes (10 cases), but the most improvement came from the updates to exon-intron boundaries of existing gene models indicating that the new gene annotation improved the overall utility of this resource.

The isoform-dependent (Cuffdiff) and-independent (MATS) analyses identified only 37 common gene models. Even though Cuffdiff revealed fewer events as compared to the MATS analysis, it did identify 100 genes with differential isoform usage that were not included in the set of genes with multiple isoforms. This suggests that differential isoform expression analysis profits from prior isoform annotation, however, should not only rely on existing isoforms.

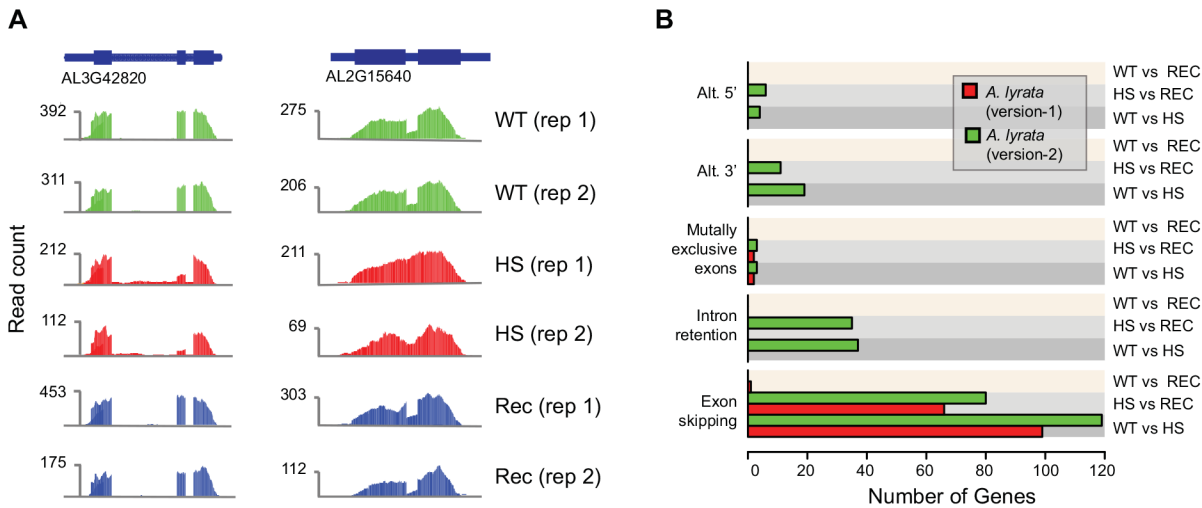


Fig 4. Heat stress induces alternative splicing events. (A) Examples of differentially expressed isoforms in response to heat stress in *A. lyrata*. AL3G42820 expresses a second isoform that lacks the middle exon in heat-treated samples (HS). Transcripts from wild-type (WT) and recovery (REC) samples contain all three exons. AL2G15640 retains an intron in response to heat stress (HS) while wild-type (WT) and recovery (REC) samples show partial intron splicing. (B) Number of differential splicing events, including alternative 5' and 3' splice sites, mutually exclusive exons, intron retention, and exon skipping events identified with MATs based on version-1 and version-2 annotations.

doi:10.1371/journal.pone.0137391.g004

Availability of the annotation and gene naming conventions

The version-2 annotation can be found in [S2 Dataset](#). The gene identifiers have been updated following the annotation principles applied to *A. thaliana* [29]. Gene numbering follows the physical order of genes on chromosomes, where each gene is named “AL” followed by scaffold number, then a “G” (for the first 8 scaffolds corresponding to the eight chromosomes) or “U” (for unanchored scaffolds) and finally a unique number incremented by 10, to leave flexibility for genes that were missed in this annotation. Genes that were removed from version-1 can be found in [S3 Dataset](#). A mapping of the gene model identifiers of version-1 to version-2 can be found in [S4 Dataset](#).

Conclusions

The updated annotation includes 31,132 gene models with 35,805 transcripts. We also reported 1,304 gene models that were erroneously split or merged in the previous annotation. Validation of these models strongly supported our updates highlighting the importance of employing species-specific RNA-seq data for annotating genomes.

We also provided a first annotation of alternative splicing events in *A. lyrata*. Using RNA-seq samples for a heat stress experiment we demonstrated the improved utility of the version-2 annotation for differential isoform expression studies. This revised genome annotation advances the reference sequence of *A. lyrata* as a community resource for comparative and functional studies.

Material and Methods

Plant material

Arabidopsis lyrata subsp. *lyrata* MN47 plants were grown in soil under long day conditions (16 hours light, 21°C: 8 hours dark, 16°C). Vegetative rosettes and dissected shoot apices of three week old plants and entire inflorescences of flowering plants were harvested as mock treated

samples. For heat stress and recovery treatments we incubated three week old plants at 37°C for 6 hours or for an additional 48 hours at 21°C, respectively. Cold stressed samples were treated as described [6].

Nucleic acids isolation and RNA-seq library preparation

DNA was isolated using Nucleon Phytopure kit (GE Healthcare). For total RNA isolation, samples were flash frozen in liquid nitrogen and used with Qiagen RNeasy[®] Plant Mini Kit, including an on-column DNase I digestion. Total RNA integrity was confirmed on the Agilent BioAnalyzer. Barcoded libraries were constructed using the Illumina TruSeq RNA kit with average of 1 µg of total RNA as starting material. The manufacturer's protocol was precisely followed with one exception in the cold-treated samples where 12 PCR cycles were used instead of the recommended 15. The library quality was monitored on a Bioanalyzer 2100 (Agilent) and the libraries were sequenced as 100-bp single end reads using Illumina sequencing.

RNA-seq read mapping and gene predictions

RNA-seq data was mapped to the *A. lyrata* reference genome assembly [1] using Bowtie v1.0.0 [26] and TopHat v2.0.10 [27]. Cufflinks v2.0.2 [22] was used for *de novo* transcript identification in all tissues separately. Cuffmerge (from the Cufflinks suite) was used to merge transcript annotation files obtained for three tissues separately. In addition, all short reads were aligned to the reference assembly of *A. lyrata* using BLAT v.34 [34] to generate an evidence file for guided gene prediction using Augustus v3.0.0. *A. lyrata* specific configuration file was generated using the version-1 annotation. To estimate agreement between Augustus and version-1 gene models, gene models with $\geq 30\%$ overlap (in respect to the shorter gene model) were considered. Gene models supported by five or more RNA-seq reads were considered as expressed irrespective of gene length.

To identify cases where wrong gene models were introduced in version-2, we first compared version-2 proteins (23,181 comparable proteins) with corresponding version-1 proteins. A total of 1,037 proteins were identified as outliers, where protein length difference was outside the range of \pm standard deviation of the distribution of length differences. For these cases version-1 and version-2 protein sequences were further compared against the proteins of their orthologs in *A. thaliana* [29] and *C. rubella* [4]. If both orthologs were more similar in length to the protein of version-1, the respective version-2 gene model was replaced with version-1.

Ortholog identification

Orthologous gene identification for both version-1 and version-2 was done separately at protein level using reciprocal best hits using blastall v2.2.25 [35] and an e-value cutoff 0.001 among five Brassicaceae species.

Identification of TE genes in version-2

Version-2 gene models harboring complete TEs [14] within their coding regions or were entirely spanned by a TE were annotated as "TE coding genes". In addition 3,909 *A. thaliana* TE genes [25] and TIGR Brassicaceae specific repeat database [36] were used to identify TE genes using blastn v2.2.25 [35].

cDNA preparation and PCR

Plants were grown on soil under long day conditions until the five-leaf stage reached after approximately three weeks. cDNA samples were prepared from 1 µg total RNA of mock-

treated rosettes using RevertAid First Strand cDNA Synthesis Kit with oligo d(T) primers (Thermo Scientific). Reverse transcriptase minus samples were processed in the same way without enzyme addition. PCR reactions were done in an Eppendorf thermal cycler using a standard program and the products were visualized on agarose gels stained with ethidium bromide. The PCR primer sequences can be found in [S5 Dataset](#).

Differential gene expression and alternative splicing

Cufflinks [22] was used to calculate differential gene expression level (FPKM) with p-value < 0.01 and log₂-fold change difference of more than 2. MATS [33] was used to investigate differential splicing events with over 0.01% splicing difference at a p-value < 0.01 and a false discovery rate of less than 1%. To control for false positives, genes with 10,000 fold or more expression difference were excluded.

Supporting Information

S1 Dataset. Supplementary figures.

(DOCX)

S2 Dataset. General feature formatted (GFF) file describing version-2 annotation.

(ZIP)

S3 Dataset. GFF file describing genes that were removed from version-1.

(ZIP)

S4 Dataset. Table describing the mapping of version-1 to version-2 gene models.

(XLSX)

S5 Dataset. Primer information for gene model validation.

(XLSX)

Acknowledgments

We thank Geo James Velikkakam for helpful discussion on gene annotation tools, J. de Meaux for seeds of *A. lyrata*, B. Eilts and R. Gentges for technical support, A. Platts, S. Wright and M. Blanchette for insights in their gene annotation and annotation of conserved non-coding sequences, and M. Koornneef for critical reading of the manuscript.

Author Contributions

Conceived and designed the experiments: KS AP. Performed the experiments: AP AA. Analyzed the data: VR. Contributed reagents/materials/analysis tools: AA BP DKS DK DW. Wrote the paper: VR KS AA AP BP DKS DK DW.

References

1. Hu TT, Pattyn P, Bakker EG, Cao J, Cheng J-F, Clark RM, et al. The arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nat Genet* 2011, May; 43(5):476–81. doi: [10.1038/ng.807](#) PMID: [21478890](#)
2. Johnston JS, Pepper AE, Hall AE, Chen ZJ, Hodnett G, Drabek J, et al. Evolution of genome size in brassicaceae. *Ann Bot* 2005; 95(1):229–35. PMID: [15596470](#)
3. Lysak MA, Koch MA, Beaulieu JM, Meister A, Leitch IJ. The dynamic ups and downs of genome size evolution in brassicaceae. *Mol Biol Evol* 2009, Jan; 26(1):85–98. doi: [10.1093/molbev/msn223](#) PMID: [18842687](#)

4. 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, Jul; 45(7):831–5. doi: [10.1038/ng.2669](https://doi.org/10.1038/ng.2669) PMID: [23749190](https://pubmed.ncbi.nlm.nih.gov/23749190/)
5. Willing E-M, Rawat V, Mandáková T, Maumus F, James GV, Nordström KJ, et al. Genome expansion of *arabis alpina* linked with retrotransposition and reduced symmetric DNA methylation. *Nature Plants* 2015; 1(2).
6. Seymour DK, Koenig D, Hagmann J, Becker C, Weigel D. Evolution of DNA methylation patterns in the brassicaceae is driven by differences in genome organization. *PLoS Genet* 2014, Nov; 10(11): e1004785. doi: [10.1371/journal.pgen.1004785](https://doi.org/10.1371/journal.pgen.1004785) PMID: [25393550](https://pubmed.ncbi.nlm.nih.gov/25393550/)
7. Schneeberger K, Ossowski S, Ott F, Klein JD, Wang X, Lanz C, et al. Reference-guided assembly of four diverse *arabidopsis thaliana* genomes. *Proc Natl Acad Sci U S A* 2011, Jun 21; 108(25):10249–54. doi: [10.1073/pnas.1107739108](https://doi.org/10.1073/pnas.1107739108) PMID: [21646520](https://pubmed.ncbi.nlm.nih.gov/21646520/)
8. Cao J, Schneeberger K, Ossowski S, Günther T, Bender S, Fitz J, et al. Whole-genome sequencing of multiple *arabidopsis thaliana* populations. *Nat Genet* 2011, Aug 28; 43(10):956–63. doi: [10.1038/ng.911](https://doi.org/10.1038/ng.911) PMID: [21874002](https://pubmed.ncbi.nlm.nih.gov/21874002/)
9. 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, Jun 23; 45(8):884–90. doi: [10.1038/ng.2678](https://doi.org/10.1038/ng.2678) PMID: [23793030](https://pubmed.ncbi.nlm.nih.gov/23793030/)
10. Wang X, Wang H, Wang J, Sun R, Wu J, Liu S, et al. The genome of the mesopolyploid crop species *brassica rapa*. *Nat Genet* 2011, Aug 28; 43(10):1035–9. doi: [10.1038/ng.919](https://doi.org/10.1038/ng.919) PMID: [21873998](https://pubmed.ncbi.nlm.nih.gov/21873998/)
11. Dassanayake M, Oh D-H, Haas JS, Hernandez A, Hong H, Ali S, et al. The genome of the extremophile crucifer *thellungiella parvula*. *Nat Genet* 2011, Aug 7; 43(9):913–8. doi: [10.1038/ng.889](https://doi.org/10.1038/ng.889) PMID: [21822265](https://pubmed.ncbi.nlm.nih.gov/21822265/)
12. Wu H-J, Zhang Z, Wang J-Y, Oh D-H, Dassanayake M, Liu B, et al. Insights into salt tolerance from the genome of the *thellungiella salsuginea*. *Proc Natl Acad Sci U S A* 2012, Jul 9; 109(30):12219–24. doi: [10.1073/pnas.1209954109](https://doi.org/10.1073/pnas.1209954109) PMID: [22778405](https://pubmed.ncbi.nlm.nih.gov/22778405/)
13. Yang R, Jarvis DE, Chen H, Beilstein MA, Grimwood J, Jenkins J, et al. The reference genome of the halophytic plant *eutrema salsugineum*. *Front Plant Sci* 2013; 4:46. doi: [10.3389/fpls.2013.00046](https://doi.org/10.3389/fpls.2013.00046) PMID: [23518688](https://pubmed.ncbi.nlm.nih.gov/23518688/)
14. Haudry A, Platts AE, Vello E, Hoen DR, Leclercq M, Williamson RJ, et al. An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. *Nat Genet* 2013, Aug; 45(8):891–8. doi: [10.1038/ng.2684](https://doi.org/10.1038/ng.2684) PMID: [23817568](https://pubmed.ncbi.nlm.nih.gov/23817568/)
15. Cheng S, van den Bergh E, Zeng P, Zhong X, Xu J, Liu X, et al. The *tarenaya hassleriana* genome provides insight into reproductive trait and genome evolution of crucifers. *Plant Cell* 2013, Aug; 25(8):2813–30. doi: [10.1105/tpc.113.113480](https://doi.org/10.1105/tpc.113.113480) PMID: [23983221](https://pubmed.ncbi.nlm.nih.gov/23983221/)
16. Oh D-H, Hong H, Lee SY, Yun D-J, Bohnert HJ, Dassanayake M. Genome structures and transcriptomes signify niche adaptation for the multiple-ion-tolerant extremophyte *schrenkiella parvula*. *Plant Physiol* 2014, Apr; 164(4):2123–38. doi: [10.1104/pp.113.233551](https://doi.org/10.1104/pp.113.233551) PMID: [24563282](https://pubmed.ncbi.nlm.nih.gov/24563282/)
17. Kitashiba H, Li F, Hirakawa H, Kawanabe T, Zou Z, Hasegawa Y, et al. Draft sequences of the radish (*raphanus sativus* L.) Genome. *DNA Res* 2014, May 16; 21(5):481–90. doi: [10.1093/dnares/dsu014](https://doi.org/10.1093/dnares/dsu014) PMID: [24848699](https://pubmed.ncbi.nlm.nih.gov/24848699/)
18. Lobréaux S, Manel S, Melodelima C. Development of an *arabis alpina* genomic contig sequence data set and application to single nucleotide polymorphisms discovery. *Mol Ecol Resour* 2014, Mar; 14(2):411–8. doi: [10.1111/1755-0998.12189](https://doi.org/10.1111/1755-0998.12189) PMID: [24128264](https://pubmed.ncbi.nlm.nih.gov/24128264/)
19. 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. doi: [10.1038/ncomms4930](https://doi.org/10.1038/ncomms4930) PMID: [24852848](https://pubmed.ncbi.nlm.nih.gov/24852848/)
20. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR. Highly integrated single-base resolution maps of the epigenome in *arabidopsis*. *Cell* 2008, May 2; 133(3):523–36. doi: [10.1016/j.cell.2008.03.029](https://doi.org/10.1016/j.cell.2008.03.029) PMID: [18423832](https://pubmed.ncbi.nlm.nih.gov/18423832/)
21. Wang Z, Gerstein M, Snyder M. RNA-Seq: A revolutionary tool for transcriptomics. *Nat Rev Genet* 2009, Jan; 10(1):57–63. doi: [10.1038/nrg2484](https://doi.org/10.1038/nrg2484) PMID: [19015660](https://pubmed.ncbi.nlm.nih.gov/19015660/)
22. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 2010, May; 28(5):511–5. doi: [10.1038/nbt.1621](https://doi.org/10.1038/nbt.1621) PMID: [20436464](https://pubmed.ncbi.nlm.nih.gov/20436464/)
23. Li Z, Zhang Z, Yan P, Huang S, Fei Z, Lin K. RNA-Seq improves annotation of protein-coding genes in the cucumber genome. *BMC Genomics* 2011; 12:540. doi: [10.1186/1471-2164-12-540](https://doi.org/10.1186/1471-2164-12-540) PMID: [22047402](https://pubmed.ncbi.nlm.nih.gov/22047402/)

24. Filichkin SA, Priest HD, Givan SA, Shen R, Bryant DW, Fox SE, et al. Genome-wide mapping of alternative splicing in *Arabidopsis thaliana*. *Genome Res* 2010, Jan; 20(1):45–58. doi: [10.1101/gr.093302.109](https://doi.org/10.1101/gr.093302.109) PMID: [19858364](https://pubmed.ncbi.nlm.nih.gov/19858364/)
25. Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, et al. The *Arabidopsis* information resource (TAIR): Improved gene annotation and new tools. *Nucleic Acids Res* 2012, Jan; 40(Database issue):D1202–10. doi: [10.1093/nar/gkr1090](https://doi.org/10.1093/nar/gkr1090) PMID: [22140109](https://pubmed.ncbi.nlm.nih.gov/22140109/)
26. Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. *Nat Methods* 2012, Apr; 9(4):357–9. doi: [10.1038/nmeth.1923](https://doi.org/10.1038/nmeth.1923) PMID: [22388286](https://pubmed.ncbi.nlm.nih.gov/22388286/)
27. Trapnell C, Pachter L, Salzberg SL. TopHat: Discovering splice junctions with *RNA*-seq. *Bioinformatics* 2009, May 1; 25(9):1105–11. doi: [10.1093/bioinformatics/btp120](https://doi.org/10.1093/bioinformatics/btp120) PMID: [19289445](https://pubmed.ncbi.nlm.nih.gov/19289445/)
28. Stanke M, Waack S. Gene prediction with a hidden markov model and a new intron submodel. *Bioinformatics* 2003, Oct; 19(Suppl 2):215–25.
29. The *Arabidopsis* Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 2000; 408(6814):796–815. PMID: [11130711](https://pubmed.ncbi.nlm.nih.gov/11130711/)
30. Lowe TM, Eddy SR. TRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* 1997, Mar 1; 25(5):955–64. PMID: [9023104](https://pubmed.ncbi.nlm.nih.gov/9023104/)
31. Fahlgren N, Jogdeo S, Kasschau KD, Sullivan CM, Chapman EJ, Laubinger S, et al. MicroRNA gene evolution in *Arabidopsis lyrata* and *Arabidopsis thaliana*. *Plant Cell* 2010, Apr; 22(4):1074–89. doi: [10.1105/tpc.110.073999](https://doi.org/10.1105/tpc.110.073999) PMID: [20407027](https://pubmed.ncbi.nlm.nih.gov/20407027/)
32. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of *RNA*-seq experiments with Tophat and cufflinks. *Nat Protoc* 2012, Mar; 7(3):562–78. doi: [10.1038/nprot.2012.016](https://doi.org/10.1038/nprot.2012.016) PMID: [22383036](https://pubmed.ncbi.nlm.nih.gov/22383036/)
33. Shen S, Park JW, Huang J, Dittmar KA, Lu Z-X, Zhou Q, et al. MATS: A bayesian framework for flexible detection of differential alternative splicing from *RNA*-seq data. *Nucleic Acids Res* 2012, Apr; 40(8):e61. doi: [10.1093/nar/gkr1291](https://doi.org/10.1093/nar/gkr1291) PMID: [22266656](https://pubmed.ncbi.nlm.nih.gov/22266656/)
34. Kent WJ. BLAT—the blast-like alignment tool. *Genome Res* 2002; 12(4):656–64. PMID: [11932250](https://pubmed.ncbi.nlm.nih.gov/11932250/)
35. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990, Oct 5; 215(3):403–10. PMID: [2231712](https://pubmed.ncbi.nlm.nih.gov/2231712/)
36. Ouyang S, Buell CR. The TIGR plant repeat databases: A collective resource for the identification of repetitive sequences in plants. *Nucleic Acids Res* 2004, Jan 1; 32(Database issue):D360–3. PMID: [14681434](https://pubmed.ncbi.nlm.nih.gov/14681434/)

5. “The genetic architecture of non-additive hybrid phenotypes in *Arabidopsis thaliana* hybrids”

Seymour DK*, Chae E*, Grimm DG*, Martín Pizarro C, Vasseur F, Rakitsch B, Borgwardt K, Koenig D, Weigel D. (In revision).

*These authors contributed equally to this work

Abstract

The ubiquity of non-parental hybrid phenotypes, such as hybrid vigor and hybrid inferiority, has interested biologists for over a century and is of considerable agricultural importance. Though examples of both phenomena have been subject to intense investigation, no general model for the molecular basis of non-additive genetic variance has emerged, and prediction of hybrid phenotypes from parental information continues to be a challenge. Here, we explore the genetics of hybrid phenotype in 435 *Arabidopsis thaliana* individuals derived from intercrosses of 30 parents in a half diallel mating scheme. We find that non-additive genetic variation is a major component of genetic variation in this population, and the genetic basis of hybrid phenotype can be mapped using genome-wide association techniques. Significant loci together can explain as much as 20% of phenotypic variation in the surveyed population and include examples that have both classical dominant and overdominant effects. Our study not only illustrates the promise of genome-wide association approaches to dissect the genetic architecture underpinning hybrid performance, but we also demonstrate the contribution of classical dominance to genetic variance.

Contributions

Conceived and designed the experiments: DKS DK EC DGG KB DW. Performed the experiments: DKS EC FV CMP. Analyzed the data: DKS DGG DK. Contributed to the writing of the manuscript: DKS EC DGG DK DW.

1 **Title**2 **The Genetic Architecture of Non-additive Hybrid Phenotypes in**3 ***Arabidopsis thaliana***

4

5 **Authors**

6 Danelle K. Seymour^{1,†}, Eunyoung Chae^{1,†}, Dominik G Grimm^{2,3,†}, Carmen Martín Pizarro^{1,§},
7 François Vasseur¹, Barbara Rakitsch^{2,4}, Karsten M. Borgwardt^{2,3}, Daniel Koenig¹ and Detlef
8 Weigel^{1*}

9

10 **Affiliations**

11 ¹Department of Molecular Biology, Max Planck Institute for Developmental Biology, 72076
12 Tübingen, Germany

13 ²Machine Learning and Computational Biology Research Group, Max Planck Institute for
14 Developmental Biology and Max Planck Institute for Intelligent Systems, 72076 Tübingen,
15 Germany

16 ³Machine Learning and Computational Biology Lab, Department of Biosystems Science and
17 Engineering, ETH Zürich, 4058 Basel, Switzerland

18 ⁴European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust
19 Genome Campus, Hinxton, Cambridge, UK

20 [§]Current address: Consejo Superior de Investigaciones Científicas, Universidad de Málaga,
21 29071 Málaga, Spain

22 [†]These authors contributed equally to this work.

23 ^{*}Corresponding author, email: weigel@weigelworld.org

24

25

26

27 **Abstract**

28 The ubiquity of non-parental hybrid phenotypes, such as hybrid vigor and hybrid inferiority,
29 has interested biologists for over a century and is of considerable agricultural importance.
30 Though examples of both phenomena have been subject to intense investigation, no general
31 model for the molecular basis of non-additive genetic variance has emerged, and prediction of
32 hybrid phenotypes from parental information continues to be a challenge. Here, we explore
33 the genetics of hybrid phenotype in 435 *Arabidopsis thaliana* individuals derived from
34 intercrosses of 30 parents in a half diallel mating scheme. We find that non-additive genetic
35 variation is a major component of heritability in this population, and the genetic basis of hybrid
36 phenotype can be mapped using genome-wide association techniques. Significant loci
37 together can explain as much as 20% of phenotypic variation in the surveyed population and
38 include examples that have both classical dominant and overdominant effects. Our study not
39 only illustrates the promise of genome-wide association approaches to dissect the genetic
40 architecture underpinning hybrid performance, but we also demonstrate the contribution of
41 classical dominance to “missing heritability”.

42

43 **Author Summary**

44 Progeny of inbred parents are often more fit than their parents. Such hybrid vigor, also known
45 as heterosis, has been observed in many species, but it is most extensively studied in plants.
46 Exploiting hybrid vigor has been the focus of many agricultural breeding programs and the
47 rewards of this approach are evident all around us. In maize, for example, hybrid seed has for
48 many decades accounted for the majority of seed planted each year in North America and
49 Europe. Despite the prevalence of this phenomenon and its agricultural importance, the
50 genetic basis of heterotic traits is still unclear. We have used a large collection of first-
51 generation hybrids in *Arabidopsis thaliana* to characterize the genetics of heterosis in this
52 model plant. We have identified loci that contribute substantially to hybrid vigor and show that
53 a subset of these exhibits classical dominance, an important finding with direct implications
54 for crop improvement.

55 **Introduction**

56 The often observed phenotypic superiority of progeny relative to their parents, or heterosis, is
57 a universal phenomenon and of great importance to plant agriculture. The earliest description
58 of heterosis (also known as hybrid vigor or superiority) dates to Darwin's studies of non-self
59 fertilization in plants. He noticed that intercrossing distantly related individuals gave rise to
60 larger, more vigorous progeny [1]. Four decades later George Shull coined the term
61 "heterosis" [2] to describe Darwin's initial observations, which he and Edward East had
62 independently rediscovered in 1908 [3,4]. Heterosis has long been of interest to evolutionary
63 biologists as a potential explanation for the ubiquity of cross-fertilization in plants and animals,
64 but it is also a central component of agricultural breeding programs. The combination of
65 hybrid seed technology and inbred line improvement has driven an unprecedented
66 improvement in maize yield over the past century [5]. Despite the economic importance of
67 heterosis and its intensive investigation in a wide spectrum of species, prediction of hybrid
68 performance from parental information remains a major challenge [6].

69 In the terms of quantitative genetics, hybrid vigor (and its opposite, inferiority)
70 describes a deviation of progeny from the phenotypic mean of the parents. This means that
71 heterosis cannot be explained by the addition of the effects of contributing alleles (additive
72 genetic variance) [7]. Non-additive genetic variance (dominance variance) can result from a
73 non-linear phenotypic effect of alleles at one locus, as in the case of dominant/recessive allele
74 pairs in classical genetics, or from epistatic interactions between loci [reviewed in 8]. Three
75 non-mutually exclusive models of intralocus interactions are commonly invoked to explain
76 heterosis. The overdominance model postulates that a single mutation may be beneficial in
77 the heterozygous state [3,4,9], and it accounts for at least a few cases of heterosis
78 [10,11,12,13] and hybrid inferiority [14,15]. The dominance hypothesis suggests that genome-
79 wide complementation of many rare, weakly deleterious alleles drives hybrid superiority
80 [16,17,18], but the small effect size of individual loci would make it difficult, if not impossible,
81 to generate direct support for this hypothesis. Finally, the pseudo-overdominance model also
82 explains heterosis with the complementation of deleterious, recessive alleles, but proposes
83 that when linked in repulsion, such alleles appear overdominant. Outside of these classical

84 models, some cases of hybrid inferiority [19,20,21,22] and hybrid superiority
85 [23,24,25,26,27,28,29,30] have been linked to epistatic interactions between parental alleles.
86 The availability of completely homozygous natural collections has made the model
87 plant *Arabidopsis thaliana* an excellent subject for studies of natural variation. These
88 collections, in combination with whole genome re-sequencing data, enable replicated
89 association mapping studies across varied environmental conditions. The drawback of highly
90 inbred lines is that the contribution of dominance to phenotype cannot be assessed.
91 *Arabidopsis thaliana* outcrossing rates of over 10% have been reported in the field,
92 suggesting that dominance may contribute to phenotypic variation in natural populations [31].
93 Here we explore the magnitude of non-additive genetic variation in *A. thaliana* using a half
94 diallel intercrossing scheme. This scheme was chosen because of its power to separate a
95 line's breeding value (additive contribution) from its performance in a specific cross (non-
96 additive contribution) [32]. Whole genome re-sequencing information is available for all 30
97 parental accessions in our scheme [33] enabling construction of hybrid genotypes and
98 genome-wide association (GWA) mapping of hybrid phenotypes. We show that non-additive
99 phenotypes are pervasive in *A. thaliana* hybrids and that the genetic basis of such
100 phenotypes can be uncovered using a modified GWA approach in our half diallel.

101 **Results**

102 **Experimental design and phenotypic analyses**

103 A half diallel was constructed by intercrossing 30 parental accessions of *A. thaliana* (Table
104 S1). These accessions were chosen because they span much of the genetic diversity in the
105 native range of the species and their genomes have been sequenced [33]. To facilitate the
106 large number of intercrosses, male sterile lines were generated by artificial miRNA
107 knockdown of the homeotic gene *AP3*, removing the need for manual emasculation (Materials
108 and methods) [22]. Because manual crossing is known to influence trait values of the progeny
109 even when using genetically identical parents [34], we manually self crossed each parental
110 line using *AP3* knockdown females and wild-type males as controls. This crossing scheme
111 resulted in 435 hybrid genotypes and 2x30 parental genotypes (both normally and manually
112 selfed lines). These were grown in 16°C long days in a completely randomized design with 5
113 replicates per genotype (Materials and methods). Plants were phenotyped for multiple traits
114 related to flowering time (days to flowering and leaves on the main shoot at flowering) and
115 final rosette size (rosette diameter and rosette dry mass). Additionally, images were taken of
116 young rosettes (21 and 29 days after sowing) and several rosette phenotypes were extracted
117 from these images (Materials and methods).

118 We often observed differences in phenotypes of progeny from natural self-fertilization
119 and progeny from manual fertilization of *AP3* knockdown females with pollen from isogenic
120 siblings (Fig 1, Fig S1). While such differences between otherwise genetically identical
121 individuals has been reported before [34], our much larger data set demonstrates that the
122 effect is not directional, with progeny of the manual crosses not always being larger than their
123 self-fertilized siblings (Fig 1, Fig S1). Importantly, the artificial miRNA itself is not the source of
124 these differences as the presence of the transgene explains very little, if any, of the total
125 phenotypic variance (Materials and methods). Instead, discrepancy between the phenotypes
126 of these two groups of parental genotypes is likely the result of strong maternal effects.
127 Knock-down of *AP3* in the female parents greatly diminishes fruit production, potentially
128 altering resource allocation. Regardless of the mechanism, the crossing process clearly

129 influenced the phenotypes of resulting progeny. With this in mind, we only used phenotypes
130 from manually crossed parents in our analyses below.

131 To understand the genetic independence of the measured traits we estimated their
132 genetic correlation (Fig S2) (Supplementary materials). Several traits (days to flowering,
133 leaves to flowering, and dry mass) were correlated and thus shared a genetic basis (Fig S2).
134 The remaining rosette phenotypes were also correlated, but were not, or only very weakly,
135 correlated to the flowering time traits, suggesting that the genetic basis of rosette size at
136 specific time points and flowering are largely independent (Fig S2).

137 We next sought to estimate the relative contributions of additive and dominance
138 components to overall phenotypic variation in our sample. With diallel designs, one can
139 evaluate the breeding value of each parent, or its general combining ability (GCA). One can
140 also estimate the performance of specific crosses or the specific combining ability (SCA) [32].
141 The SCA is a measure of deviation from the breeding values, or the expected performance of
142 a line in a particular hybrid. Because additive and dominance genetic variance can be derived
143 from estimates of GCA and SCA, respectively, it is possible to calculate both narrow- and
144 broad-sense heritability using these designs (Materials and methods). We estimated GCA,
145 SCA, and heritabilities for each phenotype (Fig 2A, Table S2) using a linear mixed model
146 (Materials and methods). Total genetic variance (broad-sense heritability) ranged from 24% to
147 78% of the total phenotypic variance (Fig 2A). Rosette traits of younger plants estimated from
148 images seemed to have much lower broad-sense heritability than adult traits. Despite the
149 large range of broad-sense heritability estimates, the dominance variance consistently
150 accounted for 38-76% of the total genetic variance (Fig 2B). We conclude that non-additivity
151 contributed substantially to the observed hybrid phenotypes.

152 The significant contribution of dominance to overall genetic variation suggests that
153 hybrids frequently diverge from the mid-parent value, indicative of (mid-parent) heterosis or
154 inferiority. For each hybrid-parent-trait combination we estimated the discrepancy of an
155 observed hybrid phenotype from the mid-parent means (d , the dominance deviation) and its
156 relationship to half the difference between the parents (a , the additive component) [7].
157 Because of the bidirectional discrepancy between self- and manually-fertilized parental

158 genotypes, phenotypic means from parental genotypes produced by manual crosses were
159 used to estimate d (Fig S3) (Materials and methods). The ratio of d to a provides a scale-free
160 estimate of the magnitude of hybrid superiority and inferiority (Fig 3). For all traits, this ratio
161 was on average always different from zero, indicative of non-additivity (Fig 3) (Wilcoxon two-
162 sided test, $p < 0.01$ after Bonferroni correction for the number of traits). Additionally, rosette
163 phenotypes of young plants displayed best-parent heterosis (area and perimeter, day 21 and
164 29, and area growth). In these cases the median values of d/a were significantly greater than
165 1 (Wilcoxon one-sided test, $p < 0.005$ after Bonferroni correction for the number of traits). The
166 ratio of d to a illustrates that both mid- and max-parent heterosis occurred in the diallel (Fig 3);
167 our experimental system is thus ideal for understanding the factors underlying non-additive
168 phenotypes.

169

170 **Model selection, simulation of phenotypes, and power analyses**

171 Our next goal was to identify loci that were contributing to dominance and heterosis using
172 GWASs. Typically, GWASs of continuous traits search for a linear relationship between
173 genotypic class and a trait of interest. For binary phenotypes, such as those frequently used
174 in human disease case-control studies, more complex genetic models, including dominance
175 and overdominance, can be explicitly tested [35,36], but this is rarely done in the study of
176 continuous traits. We selected two linear mixed models to search for associations between
177 genotype and phenotype using FaST-LMM in the easyGWAS framework (Materials and
178 methods) [37,38]. The first model used a standard linear additive SNP encoding, where the
179 homozygous major allele was represented as "0", heterozygous as "1", and the homozygous
180 minor allele as "2"; we refer to it as the "additive model" (Supplementary materials). The
181 second model, referred to as the "overdominant model", used a non-standard SNP encoding,
182 where both homozygous classes were represented as "0" and the heterozygous genotype as
183 "1" (Supplementary materials). For both models, the genetic similarity between individuals
184 was estimated by computing the realized relationship kinship matrix using the appropriate
185 SNP encodings [39] (Supplementary materials). In addition to fitting two different models to
186 our data, we chose to search for association of variants with estimated trait means and the

187 dominance deviation d (Fig S3). By mapping the dominance deviation, or the discrepancy
188 between the observed hybrid phenotype and the expected mid-parent value, we were able to
189 remove potentially confounding additive effects, which might increase sensitivity to detect
190 non-additive loci. The three tested models are summarized in Fig S4.

191 Because our design was different from those employed in previous *A. thaliana*
192 GWASs, we used simulations to estimate the power we had to detect loci with additive or
193 dominance effects (Materials and methods). We found that the additive model was extremely
194 underpowered in this data set regardless of the variance explained, a proxy for effect size, or
195 allele frequency of the causal SNP (Fig 4A) (Materials and methods). This could be the result
196 of the correlation of such sites with population structure or of the limited genetic diversity of
197 the source population. Simulations also showed that, in contrast to the additive model, the
198 overdominant model had sufficient power to detect associations with SNPs that explained a
199 range of variances and that had different minor allele frequencies (Fig 4B) (Materials and
200 methods), emphasizing the importance of the diallel design in our study.

201

202 **Association mapping of additive and non-additive phenotypes**

203 *In silico* F_1 genotypes were constructed by combining known parental genotypes (Materials
204 and methods) [33]. Informative sites were required to have complete information, with a minor
205 allele frequency of at least 10% in the diallel (Materials and methods). Due to the limited
206 genetic diversity in the founding parents, many positions were in complete linkage
207 disequilibrium (LD) across chromosomes (Materials and methods). These as well as positions
208 in LD with ten or more additional sites were excluded (Materials and methods), leaving
209 204,753 sites segregating in the diallel population (Fig S5, Fig S6).

210 We used the three approaches described above to identify informative SNPs
211 (Materials and methods) significantly associated with each phenotype in our population (Fig
212 S4). Regardless of phenotype, no significant SNP was detected using the additive model (Fig
213 5A, S7A, S8A, S9A) after correcting for multiple testing (Bonferroni threshold, $p < 3 \times 10^{-7}$)
214 (Materials and methods), consistent with the low power observed in our simulation
215 experiments. The overdominant LMM was fitted to both the trait means and the dominance

216 deviation d . Significant SNPs were detected for four phenotypes (Figs 5 B,C; S7 B,C; S8 B,C;
217 S9 B,C) (Materials and methods), with many more associations for the dominance deviation
218 than for the simple trait means (35 vs. 5 significant sites) (Table S3). Significant SNPs
219 collapsed into nine regions; four of these were significant for multiple phenotypes (Table S4).
220 To account for multiple testing across experiments, a more stringent Bonferroni correction
221 was applied within each phenotype (division of individual significance threshold [0.05] by
222 number of experiments per phenotype [3] x number of SNPs [204,753]) (Materials and
223 methods; Tables S3 and S4). Most SNPs were retained even after correction within
224 phenotype (Table S4). Phenotypes with lower heritability, particularly rosette phenotypes of
225 young plants extracted from images, showed no association with any position in the genome.
226 We also did not find any associations with adult rosette size, despite a broad-sense
227 heritability of 0.62. In conclusion, we identified a number of genomic positions that are
228 associated with both the trait means and the dominance deviation, suggesting that within-
229 locus interactions, either dominance or overdominance, contribute significantly to non-additive
230 genetic variance.

231

232 **Significant SNPs contribute heavily to genetic variance**

233 To assess the contribution of within-locus interactions to previously undetected genetic
234 variance or "missing heritability", the variance explained by each significant SNP was
235 quantified and compared to the variance explained by all tested SNPs. Variance explained by
236 all tested SNPs was computed using a LMM that fitted the kinship matrix to the phenotype of
237 interest using a cross validation strategy (Materials and methods). The model was trained
238 with a data set consisting of 90% of the hybrids and then used to predict phenotypes in the
239 test data set, the remaining 10% of hybrids, with 1,000 repetitions. The variance explained by
240 all tested SNPs accounted for 7 to 56 % of the total genetic variance (Fig 6A) using the
241 additive encoding, while it ranged from 18 to 45% (Fig 6B) using the overdominant encoding,
242 similar to earlier estimates of dominance genetic variances (Fig 2). We conclude that our
243 strategy enabled excellent phenotypic predictions and subsequent estimation of variance in

244 our diallel, but we note that variance estimates cannot be necessarily extrapolated to other
245 genotypes [40].

246 We further estimated the contribution of the significant loci to variation in the diallel
247 (Materials and methods). For traits with significant SNPs, we found that an individual SNP
248 generally had a large marginal effect and could explain from 0.02 to 19.6% of the phenotypic
249 variance (Fig 7A). The contribution of all significant SNPs was calculated using a ridge
250 regression model, together with the cross-validation strategy described above, to account for
251 non-independence, or linkage, between significant SNPs (Materials and methods). We found
252 that significant SNPs account for a large fraction of the total genetic variance, explaining up to
253 20% of the total genetic variance for some traits (LTF d and rosette dry mass d) (Fig 7B).

254

255 **Multi-locus SNP associations**

256 In addition to single SNP tests, we searched for multi-locus associations using a network-
257 guided approach implemented in SConES [41]. This approach leverages the protein
258 interaction network of *A. thaliana* to search for SNPs that together influence a phenotype;
259 however, it does not explicitly test for epistasis between pairs of loci [41]. Associated SConES
260 SNPs likely contribute to phenotypic variance either via the sum of multi-locus additive effects
261 or allelic heterogeneity at a single locus, where multiple, unlinked SNPs in or near to a gene
262 have similar phenotypic consequences. For each phenotype investigated with SConES,
263 between 0 and 324 SNPs were linked to the trait of interest (Table S5, Table S6), explaining
264 up to 40% of the total genetic variance of d when fitting the overdominant genetic model (Fig
265 8A). Less genetic variance could be explained when fitting the overdominant model to the
266 predicted mean phenotype (Fig 8B). The genetic variance explained by SConES SNPs is not
267 necessarily independent from the variance explained by SNPs detected via traditional
268 association mapping, but in this case only a single SNP was detected with both methods
269 (Table S4, Table S6).

270

271 **Significant SNPs and historical models of heterosis**

272 Historical models of heterosis have specific predictions regarding the allele frequencies of
273 causal loci. In the dominance model, causal loci are expected to be rare in the population,
274 while the overdominance model forecasts intermediate frequencies of such loci [reviewed in
275 8]. Minor allele frequencies of all tested SNPs were estimated in the 80 re-sequenced
276 genomes, from which the 30 diallel parents were drawn [33]. Minor allele frequencies for
277 SNPs with significant phenotypic associations either based on single-site associations (Fig
278 9A) or via SConES (Fig 9B) were compared to the allele frequency distributions of all tested
279 SNPs (Materials and methods). Significant SNPs had much lower minor allele frequencies
280 than the background test sets (Fig 9A,B). Additionally, random sampling of SNPs from the
281 test sets demonstrated that median allele frequencies of the random draws were always
282 higher than that of the actual data (Permutation test, $p < 0.002$ for 1000 permutations) (Fig
283 9A,B) (Materials and methods). Because of statistical limitations in GWASs, it is important to
284 note that we were unable to query the effects of truly rare variants.

285 The significant SNP set was collapsed into nine distinct regions and the effects of
286 these regions included both overdominant and dominant types (Table S7, Fig S10). The most
287 sensitive experiment, where the dominance deviation was fitted with an overdominant model,
288 detected SNPs in most regions (Table S3, S4). Of the nine regions, four behaved dominantly
289 and three overdominantly or pseudo-overdominantly with respect to the trait mean (Table S7,
290 Fig S10). The remaining two regions also tended toward overdominant behavior, but the
291 effect was mild. If overdominant ($d/a > 1$), or underdominant ($d/a < -1$), phenotypes were, in fact,
292 the result of multiple dominant loci, then the magnitude of d should increase upon inclusion of
293 additional loci. To test this, multi-locus genotypes of dominant regions were correlated with
294 trait means. The phenotypic behavior varied by multi-locus genotype and for some
295 combinations did, in fact, exhibit a trend towards overdominance (Fig S11) suggesting that at
296 least a portion of heterotic phenotypes can be attributed to the combination of multiple,
297 unlinked dominant loci.

298 The dominance hypothesis predicts that the degree of heterosis exhibited by a hybrid
299 will correlate with genetic distance between the parental genotypes [42,43,44]. Pairwise

300 genetic distances were calculated for all hybrids across the entire genome at a variety of
301 annotated sites (intergenic, intron, synonymous sites, non-synonymous sites, etc.).
302 Regardless of annotation, correlation between genetic distance and estimates of d/a were
303 only occasionally significant (Fig S12), with the direction of such rare correlation varying by
304 phenotype. Flowering time traits (DTF and LTF) were significantly positively correlated with
305 genetic distance for several polymorphism categories (maximum Spearman rank correlation
306 coefficient < 0.16). In contrast, rosette perimeter and diameter traits were negatively
307 correlated with genetic distance (minimum Spearman rank correlation coefficient > -0.16).
308 Lack of correlation in the levels of heterosis across traits has been observed before [45] and
309 other studies have failed to detect a relationship between genetic distance and the magnitude
310 of heterosis [34,46,47,48].

311

312 **Candidate genes are associated with relevant biological processes**

313 To gain insight into the biological relevance of each association experiment, gene ontology
314 (GO) analysis was performed using the top 1,000 SNPs associated with each trait (Materials
315 and methods). Flowering time related traits were associated with long-day photoperiodism,
316 photomorphogenesis and GO terms related to post-transcriptional regulation (Table S8).
317 Growth related traits extracted from the images of young plants were associated with GO
318 terms related to energy production via oxidative phosphorylation in the mitochondria. Though
319 many of these SNPs were not significant using a Bonferroni significance cutoff, the
320 enrichments observed in GO analyses suggest that our study detected additional contributing
321 loci (Table S8).

322 We measured LD surrounding high confidence SNPs in order to identify putative
323 candidate genes (Materials and methods). The nine significant regions collapsed into eight
324 linkage blocks (Table S7). In some cases, LD decayed quickly around the significant SNPs,
325 allowing the identification of high confidence candidate genes (Fig 10, S13, Table S7). One
326 region in particular, HV1.3, which has a dominant phenotypic effect on leaf number at
327 flowering, exhibits rapid LD decay (Fig 10). This short haplotype block spans a single gene,
328 *AGAMOUS-LIKE 50 (AGL50)*, which encodes a MADS-box transcription factor. Many other

329 members of the MADS-box family play critical roles in flowering time control [49,50,51], but
330 the functions of *AGL50* and its closest paralog *AGL49*, which belong to a poorly characterized
331 clade of the MADS-box family [52], have been unknown. We thus uncovered a possible role
332 for a type I MADS-box gene using heterozygous allelic status, which despite extensive
333 functional studies of MADS-box genes had not been linked to flowering before.

334 Discussion

335 The contribution of dominance to unexplained genetic variance

336 There is often a discrepancy between the heritability of a trait and inheritance in families on
337 the one hand and the genetic variance explained by loci identified in GWASs on the other
338 hand. This discrepancy is often referred to as “missing heritability” and the potential causes
339 are a subject of ongoing, intense debate in the field of quantitative genetics [reviewed in
340 53,54,55,56]. Proposed explanations include the lack of power to detect loci of small effect
341 [reviewed in 53,57], the importance of rare variants [reviewed in 53,55], the contribution of
342 multiple different alleles at the same locus (allelic heterogeneity) [41,58,59,60], the change in
343 allelic effects across environments [56], and the interactions between or within loci [61]. We
344 leverage the power of replicable inbred lines in *A. thaliana* and a carefully chosen
345 intercrossing scheme to explore the contribution of non-additive variance to phenotypic
346 variability. Non-linear models are typically ignored in the analysis of continuous traits [53], but
347 we find that a considerable portion of genetic variance is attributable to dominance and can
348 be mapped when non-additivity is explicitly considered. This approach reveals loci that would
349 go undetected using standard models, and suggests that a portion of missing heritability is
350 likely derived from dominance.

351 Though we cannot exclude a role for epistasis, we found that a single heterozygous
352 position can contribute up to 20% of the genetic variance (Fig 7A) and that the marginal effect
353 of significant SNPs is the result of single locus classical dominance in some cases. An
354 indirect test for multi-locus effects using the network-guided SConES approach [41] suggests
355 that allelic heterogeneity or multi-locus additive effects may account for additional phenotypic
356 variation in our population, and most of the loci identified in single locus scans do not overlap
357 with significant SNPs from SConES. As neither set of SNPs can explain all of the genetic
358 variance, we hypothesize that both intra- and interlocus interactions contribute to non-
359 additivity in our population, and that the genetic basis for both contributions is largely non-
360 overlapping.

361

362 ***Arabidopsis thaliana* in the context of historical heterosis models**

363 For nearly 100 years geneticists have sought to develop a unified model explaining heterosis.
364 Three leading hypotheses of intralocus interactions have been developed. The dominance
365 hypothesis suggests that individuals in a population carry a suite of rare, slightly deleterious
366 mutations that have not yet been purged by purifying selection [16,17,18,62,63]. The efficacy
367 of selection to remove weakly deleterious mutations is reduced in small populations, and
368 correspondingly, the dominance hypothesis has generally received the most empirical support
369 from studies of inbreeding depression [64,65,66,67]. If heterosis is the reverse of inbreeding
370 depression, then the degree of heterosis should positively correlate with the genetic distance
371 between parents and causal alleles should be rare with small phenotypic effects
372 [8,42,43,44,64]. Several dominantly acting loci have been shown to contribute to heterosis
373 [29,68,69], and, more recently, heterosis associated loci in maize have been shown to be
374 enriched for deleterious mutations [70]. Although dominance remains the prevailing
375 hypothesis, some of its assumptions are not consistently supported; the correlation between
376 the degree of heterosis and the genetic distance between parents is not always evident
377 [34,46,47,48,71,72,73,74] and some loci of moderate effect sizes have been identified
378 [29,68,69].

379 The overdominance hypothesis suggests that a very small number of overdominant
380 loci with large effects explain the majority of heterotic phenotypes [3,4,9]. This alternative
381 hypothesis is extremely attractive to applied researchers because overdominant loci can be
382 easily integrated into breeding programs. A number of studies have identified overdominant
383 QTL associated with hybrid vigor [23,24,69,73,75,76], but molecular identification of casual
384 variants is rare. Though a few cases of truly overdominant loci have been confirmed
385 [10,11,12,13], in one example, fine mapping of overdominant QTL has separated a single
386 overdominant locus into multiple, dominant loci acting in repulsion [77]; a situation called
387 pseudo-overdominance, which represents the third common hypothesis for heterosis.

388 Our experiments focus on two groups of traits related to fitness in *A. thaliana*,
389 flowering time and plant growth/size. We find evidence for non-additive genetic variance in
390 both types of traits, and, excluding DTF, all traits have median values of d greater than 0

391 consistent with mid-parent heterosis (Fig 3). The shift towards positive values of d is stronger
392 for early growth traits, suggesting that the genetic architectures of hybrid traits, and their
393 relationship to the general models of heterosis, may differ. Indeed, heritability is much higher
394 for flowering traits, and all but one locus from our GWASs control flowering time.

395 Our data are a poor fit for the dominance model, as we find overdominant and
396 dominant loci of medium to large effect. The variants at these loci, while segregating at lower
397 frequency than background SNPs, do not classify as rare by population genetic standards.
398 Furthermore, there is only a weak positive correlation of non-additivity with genetic distance
399 for most traits, and the strongest evidence for any relationship is a negative correlation with
400 rosette diameter (Fig S12). It is important to note that most of our observations are derived
401 from an analysis of flowering time. It is possible that the small number of causal loci identified
402 for early growth related traits is indicative of control by many small effect loci or greater
403 environmental variance in these traits as indicated by their lower heritability. Still, one locus
404 significant for growth could be identified, SNPs that best associate with growth are not
405 randomly distributed across GO annotations (Table S8), and the SConES network approach
406 identifies additional SNPs that explain a considerable fraction of growth related phenotypic
407 variance (Fig 8). Together these results suggest that heterosis for growth in *A. thaliana* may
408 be attributable to loci of smaller effect that can be detected in an appropriately powered
409 experiment.

410 Several previous studies of heterosis using controlled crosses in *A. thaliana* have
411 identified loci that exhibit all possible modes of gene action, including additive, dominant, and
412 epistatic interactions [26,27,28,78,79,80]. Lack of support for the major heterosis hypotheses
413 is also evident in crop species, particularly in maize and rice. Maize is a classical model for
414 investigating heterosis, and this outcrossing species is cited frequently as supporting the
415 dominance model of heterosis [18], even though several overdominant QTL have been
416 identified [29,30,73]. Rice, in contrast, has been touted as a model crop system that supports
417 the overdominance hypothesis, even though all three modes of gene action have been
418 uncovered in this predominantly selfing species as well [23,24,29,68,69]. The dichotomy
419 between these two model crop species has been attributed to their alternative mating

420 strategies; the large effect loci that we have found in *A. thaliana*, a selfing species, support
421 this argument.

422 That both types of single-locus genetic interactions can underlie heterotic phenotypes
423 in multiple species suggests that both single gene dominance and overdominance truly occur,
424 or that pseudo-overdominance is more prevalent than expected. Additionally, it is also
425 possible that putatively deleterious, dominant loci have pleotropic effects. If such loci
426 contribute positively to a second phenotype they could be retained during evolution for longer
427 than expected from their deleterious effects. Regardless of genetic behavior, the existence of
428 large-effect, Mendelian loci driving heterosis is a considerable boon to plant breeding
429 programs where they could easily be integrated into elite material.

430 **Materials and Methods**

431 **Generation of plant material**

432 A half diallel was constructed by manually intercrossing 30 inbred strains of *A. thaliana* [33],
433 facilitated by male sterility induced by an artificial miRNA targeting the homeotic genes *AP3*
434 [22]. In addition to the 435 hybrid combinations generated with this method, 30 manual self
435 crosses of the parental strains were performed using the same strategy. This ensured that the
436 maternal environments of the hybrid and parental genotypes were equivalent. A list of hybrid
437 and parental genotypes used in this study can be found in Table S1.

438

439 **Experimental design**

440 In total, 5 replicates of 495 genotypes were surveyed in this experiment (435 hybrid
441 genotypes, 30 parental genotypes from manual crosses, and 30 self-fertilized parental
442 genotypes). Five unsterilized seeds for each replicate were aliquoted into 1.5 ml tubes with
443 500 μ l of ddH₂O. Seeds were stratified in the dark at 4°C for 10 days. After stratification seeds
444 were sown into soil (CL T Topferde, www.einheitserde.de) pots in a completely randomized
445 design. Flats were covered with humidity domes and placed into 16°C growth chambers
446 under long day conditions (16 hrs light: 8 hrs dark) at a relative humidity of 65%. Light bulbs
447 were a mixture of Sylvania Cool White Deluxe to Warm White Deluxe fluorescent bulbs (4:2)
448 (<http://www.havells-sylvania.com/>). Humidity domes were removed after one week and pots
449 were manually thinned to one plant per pot. Plants were subsequently phenotyped for a
450 variety of traits: days to first open flower (days to flowering - DTF), rosette leaf count at the
451 first open flower (leaves to flowering - LTF), rosette diameter, and rosette dry mass. Once the
452 plants had produced ~10 siliques, the plants were sacrificed. At this point, the rosette
453 diameter was measured, the rosettes were placed into paper bags and dried at 80°C for 24
454 hours. After the rosettes were completely desiccated their weight was measured and
455 recorded. Additionally, images of each tray were taken at days 21 and 29. From these images
456 the following measurements were extracted using a custom imageJ [81] macro: area (day 21
457 and 29), perimeter (day 21 and 29), area growth (day 29 - day 21 / 8), and perimeter growth
458 (day 29 - day 21 / 8). In summary, the macro automatically segmented the images by

459 removing the background and returned rosette area and perimeter values in pixels for each
460 plant. Since the maternal plants were hemizygous for the artificial miRNA targeting *AP3*,
461 progeny derived from these crosses were segregating for the transgene. Plants that carry the
462 transgene were easily identified based on their floral and fruit morphology. To ensure that the
463 transgene did not alter the measured phenotypes, we recorded the artificial miRNA status of
464 each plant for use as a covariate in later analyses. Additionally, the dates that the plants were
465 sacrificed for rosette measurements were recorded for use as a potential covariate.

466

467 **Handling of missing data and data normalization**

468 Overall, germination rates were high in this experiment. Out of 495 surveyed genotypes, only
469 9 completely failed to germinate (7 hybrids and 2 manually selfed parents) and these lines
470 were excluded from further analyses (Table S1). Of the remaining lines, 98% of plants
471 germinated. Most germination failures only occurred in a single replicate (Fig S14). In these
472 cases (58 in total), the missing phenotypes were imputed as the mean of the phenotyped
473 replicates for each genotype. After exclusion of genotypes with failed germination and
474 imputation of the remaining missing data, each phenotype was Box-Cox transformed to
475 improve the normality of the data (Fig S15).

476

477 **Estimation of GCA, SCA, and heritability**

478 A traditional ANOVA approach was not appropriate for our data because there are some
479 missing data [32]. Instead, variance components were estimated using a linear mixed model
480 implemented in SAS using a Restricted Maximum Likelihood (REML) estimation method [82].
481 The SAS code is available in Text S1 and is a modified version of the code available on Fikret
482 Isik's webpage
483 ([http://www4.ncsu.edu/~fisik/Analysis%20of%20Diallel%20Progeny%20Test%20with%20SAS](http://www4.ncsu.edu/~fisik/Analysis%20of%20Diallel%20Progeny%20Test%20with%20SAS.pdf)
484 .pdf). Only hybrid genotypes were used. The following linear mixed model was fitted to the
485 transformed data:

$$Y_{jkl} = \mu + G_j + G_k + S_{jk} + E_{jkl}.$$

486 Here, Y_{jkl} is the l -th phenotypic observation for the jk -th cross, μ is the overall mean, G_j or
 487 G_k is the random general combining ability (GCA) of the j -th female or the k -th male, S_{jk} is the
 488 random specific combining ability (SCA) of the j -th female and the k -th male, and E_{jkl} is the
 489 error term. All terms were expected to be normally distributed. This model can also be written
 490 in matrix format:

$$491 \quad \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}.$$

492 Here, \mathbf{y} is a vector of observations, $\boldsymbol{\beta}$ is a vector of the fixed effects parameter (overall mean),
 493 $\boldsymbol{\gamma}$ is the vector of random effects parameters (GCA and SCA), $\boldsymbol{\varepsilon}$ is the random error vector, \mathbf{X}
 494 is the known design matrix for the fixed effects, and \mathbf{Z} is the known design matrix for the
 495 random effects. In SAS, the \mathbf{Z} design matrix was constructed by hand using PROC IML to
 496 associate each individual with its respective parents. Next, PROC MIXED was run on the data
 497 using the above model. Variance components and covariances of variance components were
 498 extracted from the model and used to calculate both broad- and narrow-sense heritability (as
 499 well as their standard errors). Since our parents were not derived from a randomly mated
 500 population, the additive (σ_A^2) and dominance (σ_D^2) genetic variance and the total phenotypic
 501 variance (σ_P^2) in our data were as follows [32]:

$$\begin{aligned} \sigma_A^2 &= 2\sigma_{GCA}^2, \\ \sigma_D^2 &= \sigma_{SCA}^2, \\ \sigma_P^2 &= 2\sigma_{GCA}^2 + \sigma_{SCA}^2 + \sigma_{Error}^2. \end{aligned}$$

502 Both narrow- (h_n^2) and broad-sense (H_b^2) heritabilities were calculated from these values [32]:

$$\begin{aligned} H_b^2 &= \frac{\sigma_A^2 + \sigma_D^2}{\sigma_P^2}, \\ h_n^2 &= \frac{\sigma_A^2}{\sigma_P^2}. \end{aligned}$$

503

504 **Estimation of mean genotypic values**

505 A linear mixed model was fitted to each Box-Cox transformed phenotype using the package
 506 lme4 in the R statistical framework [83]. For each phenotype, the following model was fitted:

$$Y_{jkl} = G_{jk} + A_{jk} + E_{jkl}$$

507 where G_{jk} is the random genotypic effect of the j -th female and the k -th male, A_{jk} is the
 508 random effect of the *amiR AP3* transgene on the hybrid cross of j -th female and the k -th male,
 509 and E_{jkl} is the error term. For each phenotype, the above model was fitted with and without
 510 the transgene variable and the significance of this term was tested. In a few cases the
 511 transgene term was not significant and was subsequently removed from the model (DTF,
 512 LTF, and Dry mass). In the remaining cases, the transgene explained only 0.02-2.18% of the
 513 total phenotypic variance. After model fitting, the coefficients of each genotype were extracted
 514 from the model and used for all subsequent analyses. Broad-sense heritability was also
 515 calculated from these models:

$$H_b^2 = \frac{\sigma_G^2}{\sigma_P^2}$$

516 Here, σ_G^2 is the variance due to the genotype term (G_{jk}) and σ_P^2 is the total phenotypic
 517 variance. The broad-sense heritability estimates were comparable to those derived from SAS
 518 (Fig 2).

519

520 **Calculation of the dominance deviation**

521 The predicted genotypic values of hybrid and manually selfed parental genotypes were
 522 extracted from the linear model. Using these values standard quantitative genetic
 523 components of phenotype were calculated [7]. The dominance deviation d was calculated as
 524 the distance of the hybrid phenotype from the mid-parent value, or mean of the two parental
 525 genotypes. Two of the 30 manually selfed parental genotypes did not germinate and as a
 526 result the dominance deviation could not be calculated for hybrids generated from these two
 527 parents (Bak-2 and ICE61) (Table S1). For subsequent analyses only 372 hybrid genotypes
 528 were used.

529

530 **Generation of F₁ genotypes and SNP filtering**

531 *In silico* F₁ genotypes were constructed from parental genome sequences (TAIR10) [33].
 532 Before generation of hybrid genotypes, parental genotypes were filtered to remove 1) all sites
 533 that lacked complete information, 2) all sites that were not polymorphic, 3) all triallelic sites

534 (with respect to the reference), and 4) all singletons. 723,403 SNPs remained after filtering.
535 Because the parental genotypes are few, there is extensive long-distance LD between sites.
536 To remove such sites we first encoded all 723,402 SNPs using the standard additive 0,1,2
537 encoding, where 0 is the major, 1 the heterozygous, and 2 the minor allele. After encoding,
538 75,346 SNPs are only observed once within this population. We then created categories for
539 how often a specific SNP pattern across all individuals was observed within our dataset
540 ("Pattern occurrence"). These categories ranged from 2 to 7,364. For example, a SNP is
541 located in "category 2" if this SNP shares the same pattern with exactly one other SNP in the
542 genome and into "category 1,000" if the SNP in question shares the same pattern with exactly
543 999 other SNPs. In Fig S5 the cumulative number of SNPs for all categories is plotted. We
544 observe that 32.97% of all our SNPs fall into the categories 1-10, which includes all distinct
545 SNPs plus the number of SNPs for each of the categories from 2 to 10. Approximately 38% of
546 all SNPs fall into the categories 100 – 7,364. Next we evaluated whether SNPs with shared
547 patterns were located on the same chromosome or distributed across multiple chromosomes.
548 Fig S6 shows the distribution of SNPs across chromosomes for categories 2-20. For the final
549 SNP set, we allowed SNPs to share their pattern with up to 9 other positions (categories 1-
550 10), but we removed all sites that exhibited complete long distance LD across chromosomes.
551 The final data set consisted of 204,753 SNPs and these sites were used for all association
552 mapping experiments.

553

554 **Genome wide association mapping (additive model)**

555 All GWASs were conducted using the easyGWAS framework [37]. We used a local copy of
556 easyGWAS and custom C/C++ and Python implementations of the FaSTLMM [38] algorithm.
557 For the additive model, the homozygous major allele is encoded with "0", the heterozygous
558 genotype with "1" and the homozygous minor allele with "2". The genetic similarity between all
559 genotypes was estimated by computing the realized relationship kinship matrix [39] on the
560 additively encoded genotype data. This kinship matrix was used in the FaSTLMM model to
561 account for confounding due to population stratification and cryptic relatedness. The additive
562 model was only run on the predicted phenotypic values. Genomic control (GC) values were

563 computed to assess the degree of inflated test statistics [84]. GC is measuring the deviation
 564 of the observed median test statistics from the expected one. GC values larger than 1 are
 565 indicative of inflated p-values, whereas values smaller than one are indicative of deflated p-
 566 values. GC values for each experiment can be found in Table S9 and QQ plots for
 567 phenotypes with significant SNPs can be found in Fig S16.

568

569 **Genome-wide association mapping (overdominant model)**

570 We conducted GWAS with an overdominant genotype encoding, where both the homozygous
 571 minor and homozygous major alleles are encoded as “0” and the heterozygous genotype with
 572 “1”. The kinship matrix was computed on the overdominantly encoded data. Using the
 573 overdominant encoding, GWA mapping was performed on both the predicted phenotypic
 574 values of the hybrids as well as the dominance deviation of each strain.

575

576 **Multiple hypothesis testing correction**

577 To account for multiple hypothesis testing, we used a conservative 5% Bonferroni threshold of
 578 $0.05/(\text{Number of tested SNPs [204,753]})$ ($p < 2.4 \times 10^{-7}$). This correction was performed within
 579 each experiment and significant results are reported in Table S3 and S4. Additionally, we
 580 performed an even more stringent correction by accounting for the number of experiments per
 581 phenotype (3). In this case the Bonferroni threshold was equal to $2.4 \times 10^{-7}/3 = 8.1 \times 10^{-8}$. The
 582 results from this test correction are reported in Table S3 and S4.

583

584 **Estimation of variance explained by all SNPs**

585 We computed how much of the phenotypic variance could be attributed to the genetic
 586 contribution (random effect) using a cross-validation approach. We generated 1000 randomly
 587 drawn training sets (containing 90% of all hybrid genotypes) and testing sets (remaining 10%
 588 of genotypes). We then trained the LMM using only the kinship matrix (random effect) on the
 589 training data and subsequently predicted the phenotype \hat{y} of the remaining testing set.

590 Predictions were obtained as follows:

$$\hat{y} = \mathbf{C}_{test}\tilde{\mathbf{B}} + \mathbf{K}_{test}(\mathbf{K}_{train} + \delta\mathbf{I})^{-1}(\mathbf{y}_{train} - \mathbf{C}_{train}\tilde{\mathbf{B}}),$$

591 where \mathbf{C} is a vector of ones (or different covariates if given), \mathbf{K} is the kinship matrix, and $\tilde{\boldsymbol{\beta}}$ and
 592 $\tilde{\delta}$ are the estimated parameters from the training step of the LMM. We then computed
 593 variance explained as follows:

$$v(\mathbf{y}_{test}, \hat{\mathbf{y}}) = 1 - \frac{\text{var}(\mathbf{y}_{test} - \hat{\mathbf{y}})}{\text{var}(\mathbf{y}_{test})},$$

594 where $\text{var}()$ is the variance. Note that this measure might be negative and in such cases the
 595 phenotypic mean would provide a better fit than the actual trained model. Results were
 596 averaged across all 1000 training sets.

597

598 **Estimation of variance explained by individual significant SNPs**

599 Next, we estimated the variance explained by each individual SNP. For each significantly
 600 associated SNP we generated 1,000 randomly drawn training sets (containing 90% of all
 601 hybrid genotypes) and testing sets (remaining 10% of lines). We then computed variance
 602 explained by a single SNP by fitting a linear regression:

$$v_{SNP} = 1 - \frac{\text{var}(\mathbf{y}_{test} - \tilde{\boldsymbol{\beta}}\mathbf{X})}{\text{var}(\mathbf{y}_{test})},$$

603 where $\tilde{\boldsymbol{\beta}}$ is the estimated parameter from the linear model and \mathbf{X} is the associated SNP. The
 604 parameter $\tilde{\boldsymbol{\beta}}$ is estimated as follows:

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y},$$

605 where \mathbf{X}^T is the transposed matrix \mathbf{X} . Results were average across all 1000 training sets.

606 Note that these phenotypic predictions are only relevant to the current data set and that the
 607 results (i.e. variance explained by each site) need not generalize to genotypes outside of this
 608 data set.

609

610 **Estimation of variance explained by all significant SNPs**

611 It is important to note that one cannot sum up the variance explained by all individual SNPs to
 612 obtain the variance explained by all significantly associated SNPs, because the positions are
 613 not entirely independent of one another, predominantly due to LD. To estimate the variance
 614 explained by all significantly associated SNPs, we trained a ridge regression on \mathbf{X} , where \mathbf{X}
 615 contains all significantly associated SNPs. Ridge regression includes a penalty term to

616 regularize the weight of each SNP and thus implicitly takes the relatedness between
 617 individual SNPs into account:

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y},$$

618 where λ is the penalty term. λ is optimized by performing an internal line-search for a range of
 619 λ values: $\lambda = \{1e^{-3}, 1e^{-2}, 1e^{-1}, 1, 1e^1, 1e^2, 1e^3\}$. Again, 1000 cross-validation sets were run for
 620 each model.

621

622 **Power analyses**

623 To evaluate the power of the different encoding strategies, we performed a simulation
 624 experiment in which we measured the power of each test with respect to the variance
 625 explained by the causal SNP, the minor allele frequency of the causal SNP, and the SNP
 626 encoding. All experiments were performed with both the additive and overdominant SNP
 627 encodings. We binned the tested SNPs (204,753) according to their minor allele frequency
 628 $\{0.10-0.15, \dots, 0.45-0.50\}$. As the background covariance matrix (kinship matrix) we used the
 629 realized relationship matrix based on all SNPs (204,753), applying the appropriate encoding
 630 [39]. For combinations of factors (variance explained, minor allele frequency, and SNP
 631 encoding), we first randomly chose a causal SNP with the selected minor allele frequency
 632 from our genotypic data. We simulate the phenotype as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

633 where \mathbf{X} is the causal SNP and $\boldsymbol{\beta}$ is the regression coefficient. We let the proportion of
 634 variance explained by the SNP vary between (0.05,0.10,0.20,0.40,0.60,0.80). The remaining
 635 variance is explained by Gaussian distributed noise:

$$\boldsymbol{\epsilon} \sim N(0, 1 - v_{SNP} \mathbf{I}),$$

636 where v_{SNP} is the variance explained by the focal SNP. Each combination of factors (variance
 637 explained, minor allele frequency, and SNP encoding) was repeated 1000 times. Results
 638 show the power, or 1 minus the probability of not detecting the causal SNP, averaged over all
 639 repetitions, along with the standard errors (Fig 4A,B).

640

641 **Characterization of significant peaks and identification of candidate genes**

642 For GWASs, we considered only SNPs with complete genotype information in our parental
643 panel, but this approach removes some potentially relevant polymorphism. To characterize
644 the decay of linkage disequilibrium around peaks and to develop a candidate gene list we
645 used a less stringent cutoff of 70% complete information at all sites, and identified additional
646 candidate SNPs based on linkage to significant sites using PLINK 1.9 [85]. SNPs within 200
647 kb of a significant SNP, in complete linkage disequilibrium (LD) ($r^2 = 1$), and with a minor
648 allele frequency greater than 0.1 were collapsed into the eight candidate regions listed in
649 Table S7. Two peaks on chromosome 3 were collapsed into a single large region using this
650 approach because of their physical proximity and extended LD in this region. The regional
651 information was used to develop candidate lists. Decay of LD around these peaks was
652 calculated from the reference SNP in each of the above-described regional LD blocks for up
653 to 200 kb surrounding the focal SNP (Fig 10, Fig S13).

654

655 **Gene ontology (GO) analysis**

656 For each of the GWAS experiments using the overdominant SNP encoding, the top 1000
657 most associated SNPs were compared against the complete set of tested SNPs using the
658 SNP2GO library in the R statistical computing environment [86]. Significance was established
659 using a 5% FDR threshold within each phenotype.

660

661 **Acknowledgments**

662 We thank Rebecca Schwab for discussions and reading of the manuscript and Marco
663 Todesco, Anette Habring-Müller, Marius Schneider, and Dorothee Lambert for technical
664 assistance.

665 **References**

- 666 1. Darwin C (1876) The effects of cross and self fertilisation in the vegetable kingdom.
667 London: John Murray, Albemarle Street. viii, 482 p. p.
- 668 2. Shull G (1914) Duplicate genes for capsule-form in *Bursa bursa-pastoris*. Zeits Indukt
669 Abstammungs Vererbungsl 12: 97-149.
- 670 3. Shull GH (1908) The composition of a field of maize. J Hered os-4: 296-301.
- 671 4. East EM (1908) Inbreeding in corn. Rep Conn Agric Exp Stn 1907: 419-428.
- 672 5. Duvick DN (2001) Biotechnology in the 1930s: the development of hybrid maize. Nat Rev
673 Genet 2: 69-74.
- 674 6. Riedelsheimer C, Czedik-Eysenberg A, Grieder C, Lisec J, Technow F, et al. (2012)
675 Genomic and metabolic prediction of complex heterotic traits in hybrid maize. Nat
676 Genet 44: 217-220.
- 677 7. Falconer DS, Mackay TFC (1996) Introduction to Quantitative Genetics. Harlow, Essex:
678 Addison Wesley Longman.
- 679 8. Charlesworth D, Willis JH (2009) The genetics of inbreeding depression. Nat Rev Genet
680 10: 783-796.
- 681 9. Schwartz D, Laughner WJ (1969) A molecular basis for heterosis. Science 166: 626-627.
- 682 10. Rédei GP (1962) Single locus heterosis. Zeits Vererbungsl 93: 164-&.
- 683 11. Vrebalov J, Ruezinsky D, Padmanabhan V, White R, Medrano D, et al. (2002) A MADS-
684 box gene necessary for fruit ripening at the tomato *ripening-inhibitor (rin)* locus.
685 Science 296: 343-346.
- 686 12. Krieger U, Lippman ZB, Zamir D (2010) The flowering gene *SINGLE FLOWER TRUSS*
687 drives heterosis for yield in tomato. Nat Genet 42: 459-463.
- 688 13. Guo M, Rupe MA, Wei J, Winkler C, Goncalves-Butruille M, et al. (2014) Maize *ARGOS1*
689 (*ZAR1*) transgenic alleles increase hybrid maize yield. J Exp Bot 65: 249-260.
- 690 14. Todesco M, Kim ST, Chae E, Bomblies K, Zaidem M, et al. (2014) Activation of the
691 *Arabidopsis thaliana* immune system by combinations of common *ACD6* alleles.
692 PLoS Genet 10: e1004459.
- 693 15. Smith LM, Bomblies K, Weigel D (2011) Complex evolutionary events at a tandem cluster
694 of *Arabidopsis thaliana* genes resulting in a single-locus genetic incompatibility. PLoS
695 Genet 7: e1002164.
- 696 16. Davenport CB (1908) Degeneration, albinism and inbreeding. Science 28: 454-455.
- 697 17. Bruce AB (1910) The Mendelian theory of heredity and the augmentation of vigor. Science
698 32: 627-628.
- 699 18. Jones DF (1917) Dominance of linked factors as a means of accounting for heterosis.
700 Proc Natl Acad Sci U S A 3: 310-312.
- 701 19. Bomblies K, Lempe J, Epple P, Warthmann N, Lanz C, et al. (2007) Autoimmune
702 response as a mechanism for a Dobzhansky-Muller-type incompatibility syndrome in
703 plants. PLoS Biol 5: e236.
- 704 20. Alcázar R, García AV, Parker JE, Reymond M (2009) Incremental steps toward
705 incompatibility revealed by *Arabidopsis* epistatic interactions modulating salicylic acid
706 pathway activation. Proc Natl Acad Sci USA 106: 334-339.
- 707 21. Alcázar R, García AV, Kronholm I, de Meaux J, Koornneef M, et al. (2010) Natural
708 variation at Strubbelig Receptor Kinase 3 drives immune-triggered incompatibilities
709 between *Arabidopsis thaliana* accessions. Nat Genet 42: 1135-1139.
- 710 22. Chae E, Bomblies K, Kim ST, Karelina D, Zaidem M, et al. (2014) Species-wide genetic
711 incompatibility analysis identifies immune genes as hot spots of deleterious epistasis.
712 Cell 159: 1341-1351.
- 713 23. Li ZK, Pinson SRM, Park WD, Paterson AH, Stansel JW (1997) Epistasis for three grain
714 yield components in rice (*Oryza sativa* L.). Genetics 145: 453-465.
- 715 24. Li ZK, Luo LJ, Mei HW, Wang DL, Shu QY, et al. (2001) Overdominant epistatic loci are
716 the primary genetic basis of inbreeding depression and heterosis in rice. I. Biomass
717 and grain yield. Genetics 158: 1737-1753.
- 718 25. Luo X, Fu Y, Zhang P, Wu S, Tian F, et al. (2009) Additive and over-dominant effects
719 resulting from epistatic loci are the primary genetic basis of heterosis in rice. J Integr
720 Plant Biology 51: 393-408.

- 721 26. Kusterer B, Piepho HP, Utz HF, Schon CC, Muminovic J, et al. (2007) Heterosis for
722 biomass-related traits in *Arabidopsis* investigated by quantitative trait loci analysis of
723 the triple testcross design with recombinant inbred lines. *Genetics* 177: 1839-1850.
- 724 27. Kusterer B, Muminovic J, Utz HF, Piepho HP, Barth S, et al. (2007) Analysis of a triple
725 testcross design with recombinant inbred lines reveals a significant role of epistasis in
726 heterosis for biomass-related traits in *Arabidopsis*. *Genetics* 175: 2009-2017.
- 727 28. Melchinger AE, Piepho HP, Utz HF, Muminovic J, Wegenast T, et al. (2007) Genetic basis
728 of heterosis for growth-related traits in *Arabidopsis* investigated by testcross
729 progenies of near-isogenic lines reveals a significant role of epistasis. *Genetics* 177:
730 1827-1837.
- 731 29. Garcia AA, Wang S, Melchinger AE, Zeng ZB (2008) Quantitative trait loci mapping and
732 the genetic basis of heterosis in maize and rice. *Genetics* 180: 1707-1724.
- 733 30. Guo T, Yang N, Tong H, Pan Q, Yang X, et al. (2014) Genetic basis of grain yield
734 heterosis in an "immortalized F₂" maize population. *Theor Appl Genet* 127: 2149-
735 2158.
- 736 31. Bomblies K, Yant L, Laitinen R, Kim S-T, Hollister JD, et al. (2010) Local-scale patterns of
737 genetic variability, outcrossing and spatial structure in natural stands of *Arabidopsis*
738 *thaliana*. *PLoS Genet* 6: e1000890.
- 739 32. Lynch M, Walsh B (1998) *Genetics and analysis of quantitative traits*. Sunderland, Ma.:
740 Sinauer. xvi, 980 p. p.
- 741 33. Cao J, Schneeberger K, Ossowski S, Günther T, Bender S, et al. (2011) Whole-genome
742 sequencing of multiple *Arabidopsis thaliana* populations. *Nat Genet* 43: 956-963.
- 743 34. Meyer RC, Törjék O, Becher M, Altmann T (2004) Heterosis of biomass production in
744 *Arabidopsis*. Establishment during early development. *Plant Physiol* 134: 1813-1823.
- 745 35. Sasieni PD (1997) From genotypes to genes: Doubling the sample size. *Biometrics* 53:
746 1253-1261.
- 747 36. Sladek R, Rocheleau G, Rung J, Dina C, Shen L, et al. (2007) A genome-wide association
748 study identifies novel risk loci for type 2 diabetes. *Nature* 445: 881-885.
- 749 37. Grimm D, Greshake B, Kleeberger S, Lippert C, Stegle O, et al. (2012) easyGWAS: An
750 integrated interspecies platform for performing genome-wide association studies.
751 arXiv: 1212.4788v1211.
- 752 38. Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, et al. (2011) FaST linear mixed
753 models for genome-wide association studies. *Nat Methods* 8: 833-835.
- 754 39. Hayes BJ, Visscher PM, Goddard ME (2009) Increased accuracy of artificial selection by
755 using the realized relationship matrix. *Genet Res* 91: 47-60.
- 756 40. Wray NR, Yang J, Hayes BJ, Price AL, Goddard ME, et al. (2013) Pitfalls of predicting
757 complex traits from SNPs. *Nat Rev Genet* 14: 507-515.
- 758 41. Azencott CA, Grimm D, Sugiyama M, Kawahara Y, Borgwardt KM (2013) Efficient
759 network-guided multi-locus association mapping with graph cuts. *Bioinformatics* 29:
760 i171-179.
- 761 42. Charcosset A, Lefortbuson M, Gallais A (1991) Relationship between heterosis and
762 heterozygosity at marker loci - a theoretical computation. *Theor Appl Genet* 81: 571-
763 575.
- 764 43. Charcosset A, Essioux L (1994) The effect of population structure on the relationship
765 between heterosis and heterozygosity at marker loci. *Theor Appl Genet* 89: 336-343.
- 766 44. Bernardo R (1992) Relationship between single-cross performance and molecular marker
767 heterozygosity. *Theor Appl Genet* 83: 628-634.
- 768 45. Flint-Garcia SA, Buckler ES, Tiffin P, Ersoz E, Springer NM (2009) Heterosis is prevalent
769 for multiple traits in diverse maize germplasm. *PLoS One* 4: e7433.
- 770 46. Cerna FJ, Cianzio SR, Rafalski A, Tingey S, Dyer D (1997) Relationship between seed
771 yield heterosis and molecular marker heterozygosity in soybean. *Theor Appl Genet*
772 95: 460-467.
- 773 47. Liu ZQ, Pei Y, Pu ZJ (1999) Relationship between hybrid performance and genetic
774 diversity based on RAPD markers in wheat, *Triticum aestivum* L. *Plant Breeding* 118:
775 119-123.
- 776 48. Riday H, Brummer EC, Campbell TA, Luth D, Cazcarro PM (2003) Comparisons of
777 genetic and morphological distance with heterosis between *Medicago sativa* subsp.
778 *sativa* and subsp. *falcata*. *Euphytica* 131: 37-45.

- 779 49. Smaczniak C, Immink RG, Angenent GC, Kaufmann K (2012) Developmental and
780 evolutionary diversity of plant MADS-domain factors: insights from recent studies.
781 *Development* 139: 3081-3098.
- 782 50. Posé D, Verhage L, Ott F, Yant L, Mathieu J, et al. (2013) Temperature-dependent
783 regulation of flowering by antagonistic FLM variants. *Nature* 503: 414-417.
- 784 51. Lee JH, Ryu HS, Chung KS, Pose D, Kim S, et al. (2013) Regulation of temperature-
785 responsive flowering by MADS-box transcription factor repressors. *Science* 342: 628-
786 632.
- 787 52. Parenicová L, de Folter S, Kieffer M, Horner DS, Favalli C, et al. (2003) Molecular and
788 phylogenetic analyses of the complete MADS-box transcription factor family in
789 *Arabidopsis*: new openings to the MADS world. *Plant Cell* 15: 1538-1551.
- 790 53. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, et al. (2009) Finding the
791 missing heritability of complex diseases. *Nature* 461: 747-753.
- 792 54. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, et al. (2010) Missing heritability and
793 strategies for finding the underlying causes of complex disease. *Nat Rev Genet* 11:
794 446-450.
- 795 55. Gibson G (2012) Rare and common variants: twenty arguments. *Nat Rev Genet* 13: 135-
796 145.
- 797 56. Mackay TF (2014) Epistasis and quantitative traits: using model organisms to study gene-
798 gene interactions. *Nat Rev Genet* 15: 22-33.
- 799 57. Ehrenreich IM, Torabi N, Jia Y, Kent J, Martis S, et al. (2010) Dissection of genetically
800 complex traits with extremely large pools of yeast segregants. *Nature* 464: 1039-
801 1042.
- 802 58. Sugiyama M, Azencott C-A, Grimm D, Kawahara Y, Borgwardt KM (2014) Multi-task
803 feature selection on multiple networks via maximum flows. *Proc SIAM Int Conf Data*
804 *Mining 2014*: 199-207.
- 805 59. Llinares-López F, Grimm DG, Bodenham DA, Gieraths U, Sugiyama M, et al. (2015)
806 Genome-wide detection of intervals of genetic heterogeneity associated with complex
807 traits. *Bioinformatics* 31: i240-249.
- 808 60. Sham PC, Purcell SM (2014) Statistical power and significance testing in large-scale
809 genetic studies. *Nat Rev Genet* 15: 335-346.
- 810 61. Bloom JS, Ehrenreich IM, Loo WT, Lite TL, Kruglyak L (2013) Finding the sources of
811 missing heritability in a yeast cross. *Nature* 494: 234-237.
- 812 62. Fisher RA (1930) *The Genetical Theory of Natural Selection*. Oxford: Oxford University
813 Press. 318 p.
- 814 63. Kimura M (1983) Rare variant alleles in the light of the neutral theory. *Molecular Biology*
815 *and Evolution* 1: 84-93.
- 816 64. Charlesworth D, Charlesworth B (1987) Inbreeding depression and its evolutionary
817 consequences. *Annu Rev Ecol Syst* 18: 237-268.
- 818 65. Barrett SCH, Charlesworth D (1991) Effects of a change in the level of inbreeding on the
819 genetic load. *Nature* 352: 522-524.
- 820 66. Willis JH (1992) Genetic analysis of inbreeding depression caused by chlorophyll-deficient
821 lethals in *Mimulus guttatus*. *Heredity* 69: 562-572.
- 822 67. Crow JF (1993) Mutation, mean fitness, and genetic load. *Oxford Surv Evol Biol* 9: 3-42.
- 823 68. Xiao JH, Li JM, Yuan LP, Tanksley SD (1995) Dominance is the major genetic basis of
824 heterosis in rice as revealed by QTL analysis using molecular markers. *Genetics* 140:
825 745-754.
- 826 69. Hua J, Xing Y, Wu W, Xu C, Sun X, et al. (2003) Single-locus heterotic effects and
827 dominance by dominance interactions can adequately explain the genetic basis of
828 heterosis in an elite rice hybrid. *Proc Natl Acad Sci U S A* 100: 2574-2579.
- 829 70. Mezouk S, Ross-Ibarra J (2014) The pattern and distribution of deleterious mutations in
830 maize. *G3* 4: 163-171.
- 831 71. Smith OS, Smith JSC, Bowen SL, Tenborg RA, Wall SJ (1990) Similarities among a group
832 of elite maize inbreds as measured by pedigree, F₁ grain yield, grain yield, heterosis,
833 and RFLPs. *Theor Appl Genet* 80: 833-840.
- 834 72. Barbosa AMM, Geraldi IO, Benchimol LL, Garcia AAF, Souza CL, et al. (2003)
835 Relationship of intra- and interpopulation tropical maize single cross hybrid

- 836 performance and genetic distances computed from AFLP and SSR markers.
837 *Euphytica* 130: 87-99.
- 838 73. Stuber CW, Lincoln SE, Wolff DW, Helentjaris T, Lander ES (1992) Identification of
839 genetic factors contributing to heterosis in a hybrid from two elite maize inbred lines
840 using molecular markers. *Genetics* 132: 823-839.
- 841 74. Laripe A, Mangin B, Jasson S, Combes V, Dumas F, et al. (2012) The genetic basis of
842 heterosis: multiparental quantitative trait loci mapping reveals contrasted levels of
843 apparent overdominance among traits of agronomical interest in maize (*Zea mays*
844 L.). *Genetics* 190: 795-811.
- 845 75. Pogson GH (1991) Expression of overdominance for specific activity at the
846 phosphoglucosmutase-2 locus in the Pacific oyster, *Crassostrea gigas*. *Genetics* 128:
847 133-141.
- 848 76. Mitchell-Olds T (1995) Interval mapping of viability loci causing heterosis in *Arabidopsis*.
849 *Genetics* 140: 1105-1109.
- 850 77. Graham GI, Wolff DW, Stuber CW (1997) Characterization of a yield quantitative trait
851 locus on chromosome five of maize by fine mapping. *Crop Sci* 37: 1601-1610.
- 852 78. Lisec J, Steinfath M, Meyer RC, Selbig J, Melchinger AE, et al. (2009) Identification of
853 heterotic metabolite QTL in *Arabidopsis thaliana* RIL and IL populations. *Plant J* 59:
854 777-788.
- 855 79. Meyer RC, Kusterer B, Lisec J, Steinfath M, Becher M, et al. (2010) QTL analysis of early
856 stage heterosis for biomass in *Arabidopsis*. *Theor Appl Genet* 120: 227-237.
- 857 80. Oakley CG, Ågren J, Schemske DW (2015) Heterosis and outbreeding depression in
858 crosses between natural populations of *Arabidopsis thaliana*. *Heredity* 115: 73-82.
- 859 81. Schneider CA, Rasband WS, Eliceiri KW (2012) NIH Image to ImageJ: 25 years of image
860 analysis. *Nat Methods* 9: 671-675.
- 861 82. Xiang B, Li BL (2003) Best linear unbiased prediction of clonal breeding values and
862 genetic values from full-sib mating designs. *Can J For Res* 33: 2036-2043.
- 863 83. Bates D, Maechler M, Bolker B, Walker S (2015) lme4: Linear mixed-effects models using
864 Eigen and S4. R package version 11-8.
- 865 84. Devlin B, Roeder K (1999) Genomic control for association studies. *Biometrics* 55: 997-
866 1004.
- 867 85. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, et al. (2015) Second-generation
868 PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 4: 7.
- 869 86. Szkiba D, Kapun M, von Haeseler A, Gallach M (2014) SNP2GO: functional analysis of
870 genome-wide association studies. *Genetics* 197: 285-289.

871

872 **Figure legends**

873 **Figure 1. Effect of manual- versus self-fertilization on parental phenotypes.** Predicted
874 trait means for each parental line were estimated using a linear mixed model (Materials and
875 methods), and untransformed ratios are shown.

876

877 **Figure 2. Broad- and narrow-sense heritability estimates.** (A) Broad- and narrow-sense
878 heritabilities for each phenotype. Heritabilities were calculated from estimates of GCA
879 (general combining ability) and SCA (specific combining ability), which were derived from a
880 LMM implemented in SAS (Materials and methods). Yellow circle shows the broad-sense
881 heritability re-estimated using a linear model in R (Materials and methods). (B) Dominance
882 genetic variance as a fraction of total genetic variance for all phenotypes.

883

884 **Figure 3. Distribution of non-additive phenotypes.** The ratio of the dominance deviation d
885 to the additive phenotypic component a is shown for all hybrid genotypes for each phenotype
886 (Materials and methods). Boxplot shows the median (white circle), upper and lower quartiles
887 (black box), and 1.5x the interquartile range (black lines). Values greater than 1 are
888 overdominant (best-parent heterosis) and values less than -1 are underdominant (worst-
889 parent hybrid inferiority).

890

891 **Figure 4. Power analyses of additive and overdominant models.** Power of the additive
892 (A) and overdominant (B) models was calculated for various minor allele frequencies (MAF)
893 and effect sizes. Variance explained by the focal SNP is considered a proxy for effect size
894 and varies between 5 and 80 percent of the variance (Materials and methods). Simulations
895 were performed 1,000 times for each combination of MAF and percent of explained variance.
896 Mean power of the simulations as well as the standard error of the mean are plotted. β is the
897 rate of Type II error.

898

899 **Figure 5. Genome-wide association studies of days to flowering (DTF) phenotype.**

900 Genome wide p-values of each test statistic are shown for each association experiment

901 performed on the DTF phenotype. Horizontal dashed lines correspond to 5% significance
902 thresholds for both within ($p < 2.4 \times 10^{-7}$, orange) and across phenotype ($p < 8.1 \times 10^{-8}$, blue)
903 Bonferroni corrections. Genomic controls were estimated as the deviation of the observed
904 median test statistics from expected median test statistic. (A) Additive model - mean
905 phenotype. (B) Overdominant model - mean phenotype. (C) Overdominant model -
906 dominance deviation.

907

908 **Figure 6. Phenotypic variance explained by genome-wide SNPs.** Variance explained
909 using all tested SNPs was calculated using a cross-validation approach (Materials and
910 methods). Mean variance explained and the standard error of the mean (1,000 training sets)
911 are plotted for the training (90%) and test (10%) sets. (A) Phenotypic variance of the mean
912 trait value explained by the additive model. (B) Phenotypic variance of the dominance
913 deviation explained by the overdominant model. Models that are evaluated using their own
914 training data tend to overfit, hence the values that are close to, or equal to, 1.

915

916 **Figure 7. Phenotypic variance explained by significantly associated SNPs.** Variance
917 explained by individual SNPs (A) and all significantly associated SNPs (B) was calculated
918 using a cross-validation approach (Materials and methods). (A) Boxplot shows the median
919 (white circle), upper and lower quartiles (black box), and 1.5x the interquartile range (black
920 lines). (B) Mean variance explained and the standard error of the mean (1,000 training sets)
921 are shown for the training (90%) and test (10%) sets.

922

923 **Figure 8. Phenotypic variance explained by SConES SNPs.** Variance explained using all
924 associated SNPs identified via SConES (multi-locus network-guided associations) was
925 calculated using a cross-validation approach (Materials and methods). Mean variance
926 explained and the standard error of the mean (1000 training sets) are shown for the training
927 (90%) and test (10%) sets. (A) Phenotypic variance of the dominance deviation d explained
928 by the overdominant model. (B) Phenotypic variance of the mean trait value explained by the
929 overdominant model.

930

931 **Figure 9. Minor allele frequency distributions of associated SNPs.** Minor allele frequency
932 (MAF) distributions for significant SNPs found via single-locus GWASs (A) or multi-locus
933 SConES analysis (B). MAFs of all tested SNPs (black) were calculated for an expanded panel
934 of 80 accessions. Random sampling of equal sample sizes (grey) of the test SNPs was
935 repeated 1,000 times. Median values of random draws never reached the median values of
936 the actual data (Permutation test, $p < 0.002$).

937

938 **Figure 10. Linkage disequilibrium (LD) near candidate region HV1.3.** LD is plotted for (A)
939 400 kb and (B) 20 kb surrounding region HV1.3 (Supplementary materials). Region HV1.3 (A)
940 is also shown in Fig S13. Significant SNPs associated with LTF *d* are plotted as blue circles
941 (Table S4). Gene models (TAIR10) are shown in (B) and *AGL50* (AT1G59810) is highlighted
942 in purple.

943 **Supplementary Figure Legends**

944 **Figure S1. Parental phenotypes per trait.** The mean of manual- (grey) and self-fertilized
945 (blue) parental phenotypes (untransformed) is plotted as well as the phenotypic values of
946 each individual (circles).

947

948 **Figure S2. Genetic correlation of traits.** The genetic correlation of all traits is shown.
949 Genetic correlation was calculated as the $\text{cov}_{12}/(\sigma_1\sigma_2)$ [7]. Here, cov_{12} is the covariance in
950 predicted means between each trait pair and σ_1 and σ_2 are the genetic standard deviations of
951 the line means for each trait. A trait's correlation with itself is not equal to 1 because predicted
952 line means were used instead of the measured trait value.

953

954 **Figure S3. Distribution of the dominance deviation d .** The distribution of d is plotted for
955 each phenotype. Values are expected to be distributed around zero (dashed vertical line).

956

957 **Figure S4. Schematic of GWAS experiments.** Cartoon depicting the selected SNP
958 encoding and phenotypic component for each GWA experiment. As an example, a focal SNP
959 genotype is shown as G/G, G/T, and T/T. Parental germplasm is abbreviated as P1 and P2
960 for Parent 1 and Parent 2. The focal phenotype of each GWAS experiment is marked as
961 either a green circle (the predicted trait mean of each hybrid) or as a purple line (the
962 dominance deviation d).

963

964 **Figure S5. Cumulative distribution of SNP pattern occurrence.** SNP pattern occurrence
965 was calculated as the number of positions that share a genotypic pattern (Materials and
966 methods).

967

968 **Figure S6. Chromosomal location of SNPs with shared patterns.** The number of
969 chromosomes covered by SNPs in pattern occurrence categories 2-20 are plotted.

970

971 **Figure S7. Genome-wide association studies of leaves to flowering (LTF) phenotype.**

972 Genome wide p-values of each test statistic are shown for each association experiment
973 performed on the LTF phenotype. Horizontal dashed lines correspond to 5% significance
974 thresholds for both within ($p < 2.4 \times 10^{-7}$, orange) and across ($p < 8.1 \times 10^{-8}$, blue) Bonferroni
975 corrections. Genomic controls are estimated as the deviation of the observed median test
976 statistics from expected. A) Additive model - mean phenotype, B) Overdominant model -
977 mean phenotype, C) Overdominant model - dominance deviation.

978

979 **Figure S8. Genome-wide association studies of dry mass phenotype.** Genome wide p-

980 values of each test statistic are shown for each association experiment performed on the dry
981 mass phenotype. Horizontal dashed lines correspond to 5% significance thresholds for both
982 within ($p < 2.4 \times 10^{-7}$, orange) and across ($p < 8.1 \times 10^{-8}$, blue) Bonferroni corrections. Genomic
983 controls are estimated as the deviation of the observed median test statistics from expected.
984 A) Additive model - mean phenotype, B) Overdominant model - mean phenotype, C)
985 Overdominant model - dominance deviation.

986

987 **Figure S9. Genome-wide association studies of area (day 29) phenotype.** Genome wide

988 p-values of each test statistic are shown for each association experiment performed on the
989 area (day 29) phenotype. Horizontal dashed lines correspond to 5% significance thresholds
990 for both within ($p < 2.4 \times 10^{-7}$, orange) and across ($p < 8.1 \times 10^{-8}$, blue) Bonferroni corrections.
991 Genomic controls are estimated as the deviation of the observed median test statistics from
992 expected. A) Additive model - mean phenotype, B) Overdominant model - mean phenotype,
993 C) Overdominant model - dominance deviation.

994

995 **Figure S10. Phenotype-genotype relationship of each region.** Phenotypic distribution of

996 mean phenotypes for each focal region. Only phenotypes significantly associated with each
997 region are shown, regardless of whether the association was originally detected with the
998 predicted trait mean or the dominance deviation d . The genotype of each region can
999 represent one or more SNPs (Table S4) (minor allele=1; major allele=0).

1000

1001 **Figure S11. Phenotype-multi-locus genotype relationships.** Phenotypic distribution of
1002 mean phenotypes for each multi-locus genotype. Distributions are only shown for multi-locus
1003 genotypic combinations between dominant loci. Only phenotypes significantly associated with
1004 each region are shown, regardless of whether the association was originally detected with the
1005 predicted trait mean or the dominance deviation d . The genotype of each region can
1006 represent one or more SNPs (Table S4) (minor allele=1; major allele=0).

1007

1008 **Figure S12. Correlation of d/a with genetic distance.** Pairwise genetic distance was
1009 calculated for each line for various annotation categories (Materials and methods). The
1010 spearman rank correlation coefficient was calculated for the correlation of each measure of
1011 genetic distance with the degree of heterosis (d/a). Significant correlations are indicated with *
1012 ($p < 0.05$) or ** ($p < 0.01$).

1013

1014 **Figure S13. Linkage disequilibrium decay around candidate SNPs.** Linkage
1015 disequilibrium (LD) decay (r^2) is plotted for each of the candidate regions (Materials and
1016 methods). Region HV1.3 is also shown in Fig 10. Significant SNPs detected in one of the
1017 GWA experiments (Table S4) are plotted as blue circles.

1018

1019 **Figure S14. Summary of missing phenotypic data.** Number of replicates with no
1020 phenotypic data for each genotype.

1021

1022 **Figure S15. Distribution of phenotypic values.** Distribution of raw (grey) and transformed
1023 (blue) phenotypic values for each phenotype. Phenotypes were transformed using Box-Cox
1024 (Materials and methods).

1025

1026 **Figure S16. Quantile-Quantile (Q-Q) plot for GWA experiments with significant results.**

1027 The distribution of observed p -values is plotted against the expected distribution.

1028

1029 **Supplementary Table Legends**

1030 **Table S1. Germplasm information.** The parental genotypes of each hybrid are listed along
1031 with whether the line was included in the final data set (yes=1,no=0). Lines were not included
1032 if their manually-fertilized parent did not germinate.

1033

1034 **Table S2. GCA, SCA, and heritability estimates.** Values plotted in Fig 2.

1035

1036 **Table S3. Summary of significant SNPs detected in GWASs.** Number of significant SNPs
1037 detected for each phenotype for both within-experiment and across-experiment Bonferroni
1038 correction.

1039

1040 **Table S4. Summary of significant SNPs per phenotype.** IDs (chr_position) of all
1041 significant SNPs and their significance status in each experiment.

1042

1043 **Table S5. Summary of SConES SNPs.** Number of associated SNPs detected for each
1044 phenotype.

1045

1046 **Table S6. Summary of associated SConES SNPs per phenotype.** IDs (chr_position) of all
1047 associated SNPs and their significance status in each experiment.

1048

1049 **Table S7. Summary of candidate regions.** IDs and location of all significant regions, their
1050 associated phenotypes, genetic behavior, and putative candidate genes.

1051

1052 **Table S8. Gene Ontology (GO) enrichment.** GO terms significantly associated with the top
1053 1000 associated SNPs of each phenotype are listed. Significance was established using a 5%
1054 FDR threshold within each phenotype.

1055

1056 **Table S9. Genomic control (GC) values.** Genomic control values for all experiments. Q-Q
1057 plots for experiments with significant SNPs are shown in Fig S16.

Figure 1

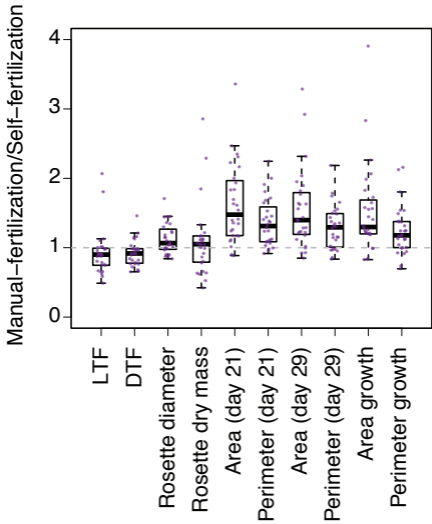
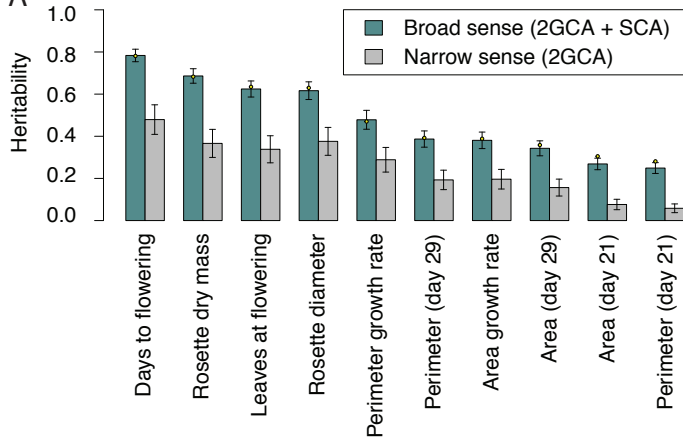


Figure 2

A



B

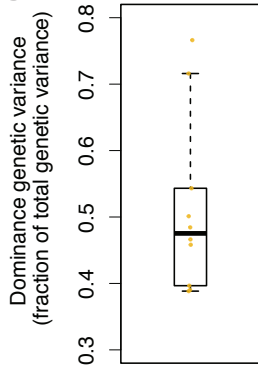


Figure 3

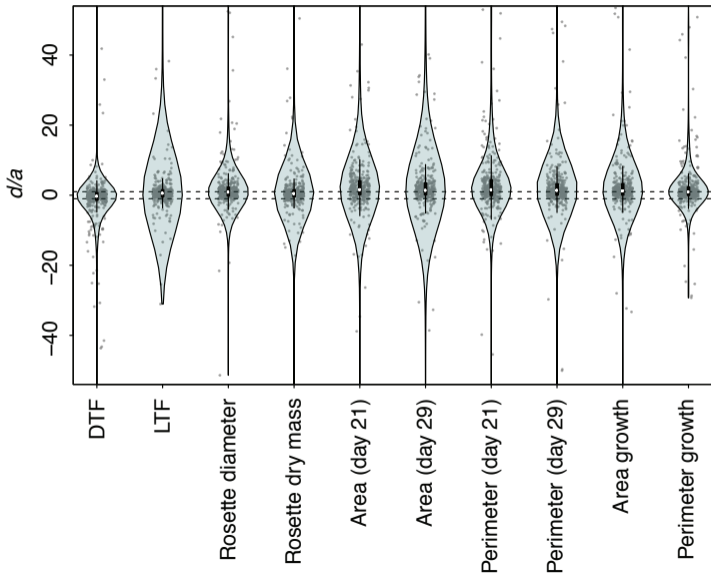


Figure 4

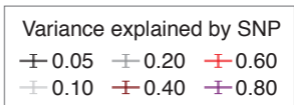
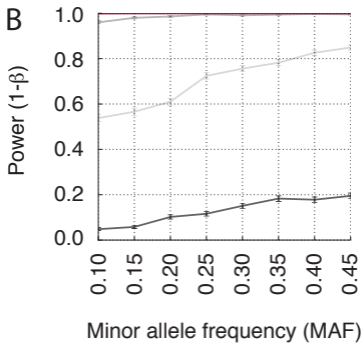
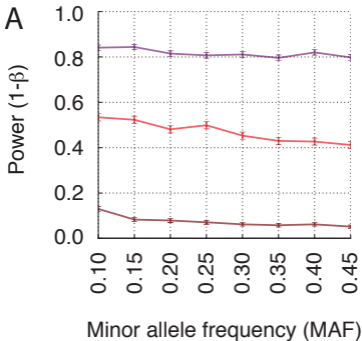


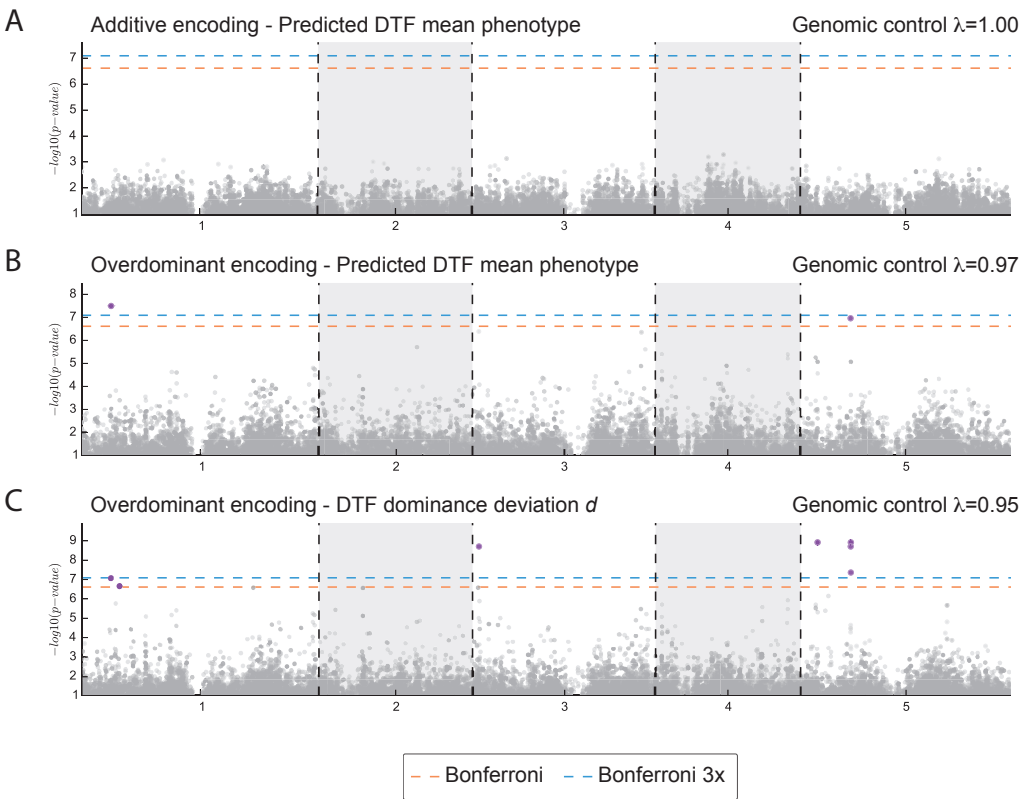
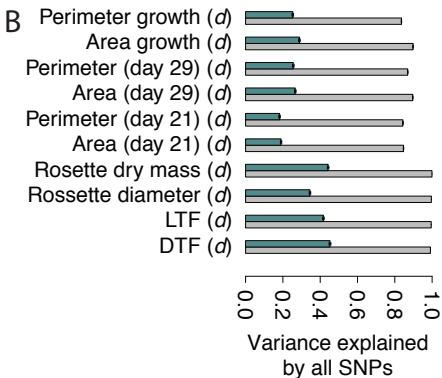
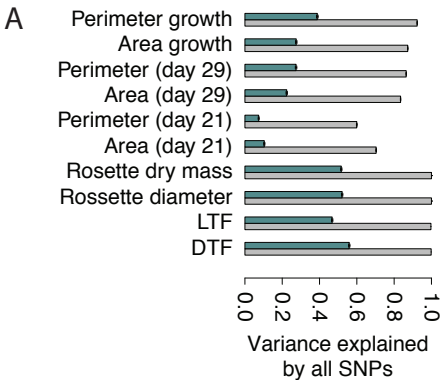
Figure 5

Figure 6



□ Training data ■ Testing data

Figure 7

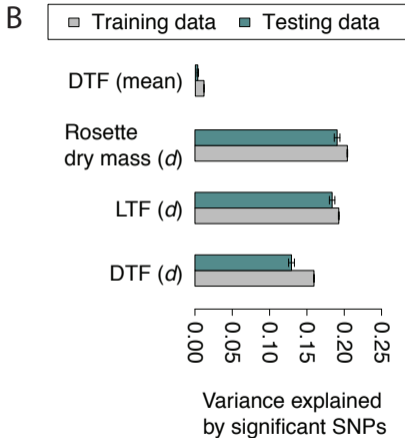
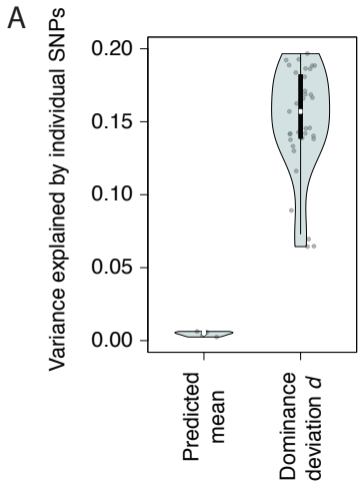


Figure 8

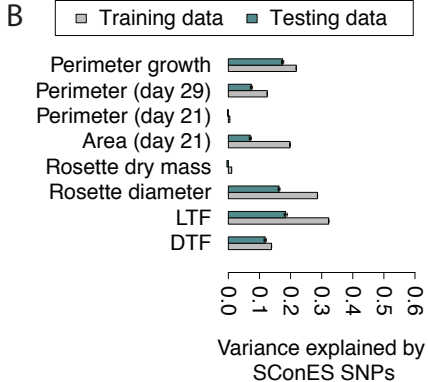
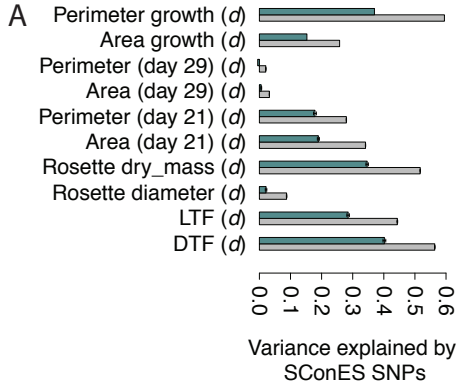
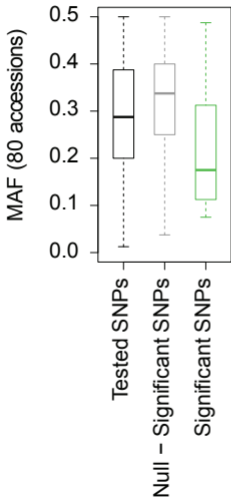


Figure 9

A



B

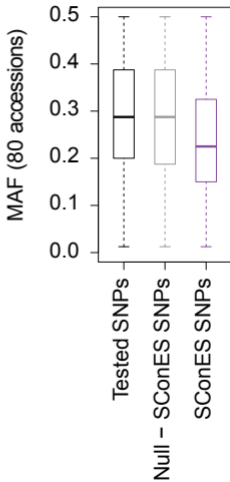


Figure 10

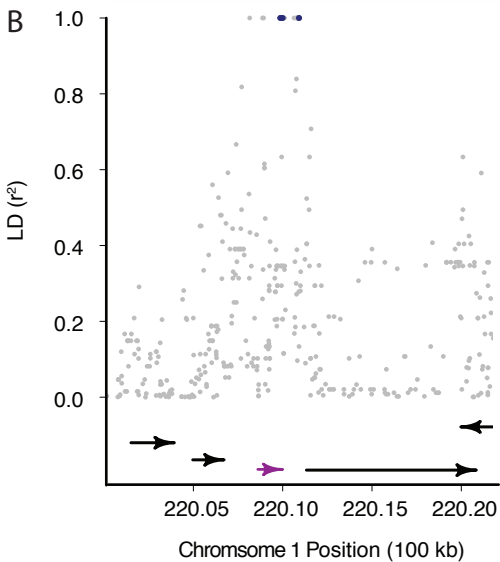
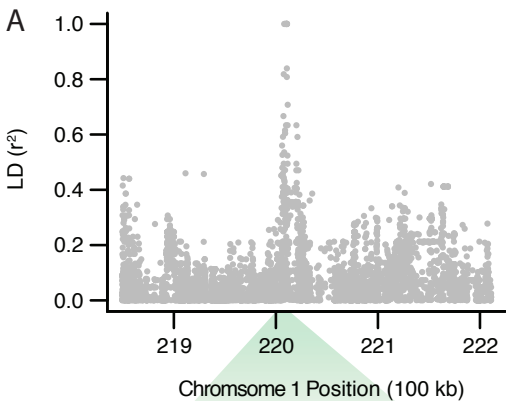


Figure S1

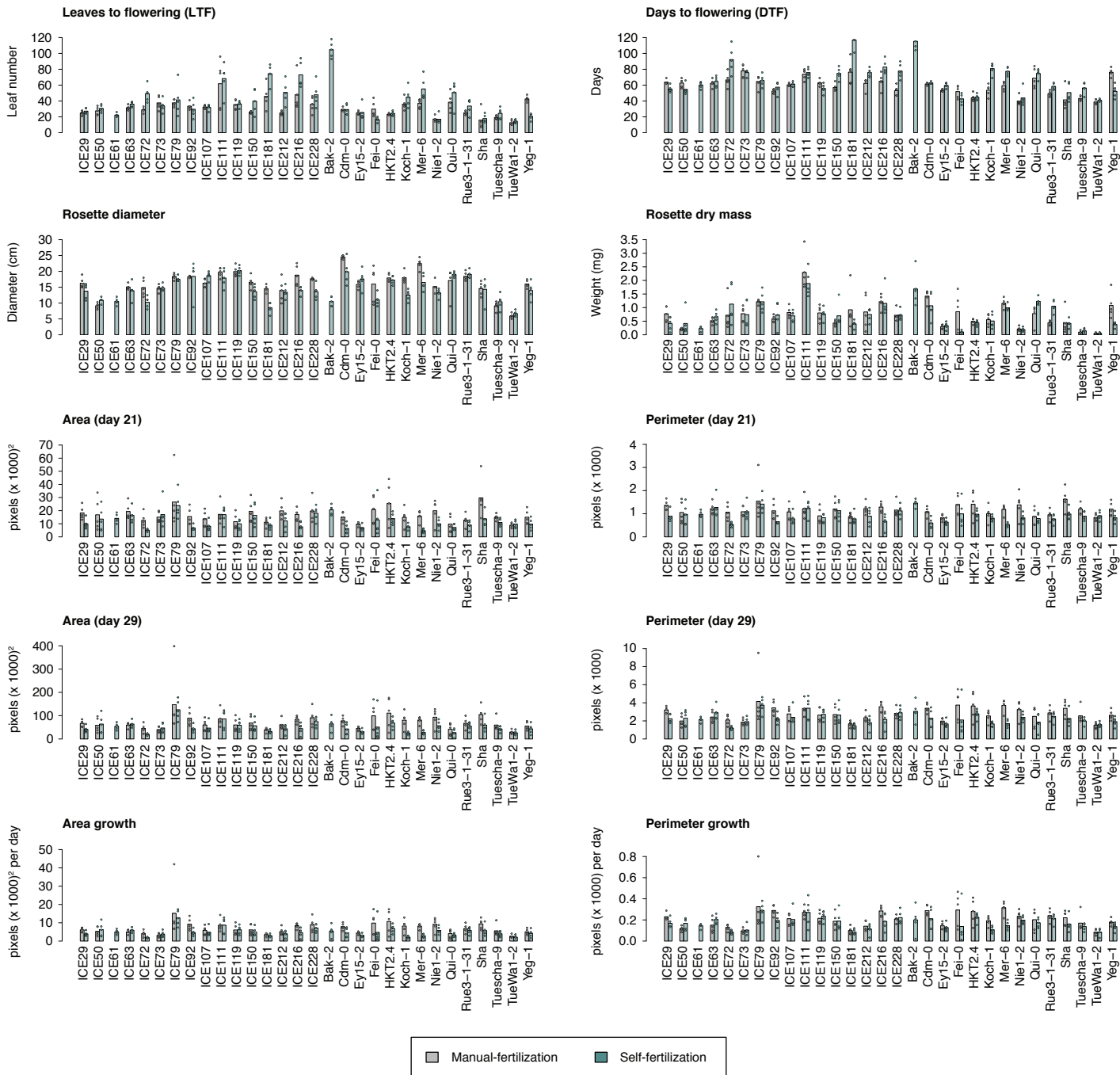


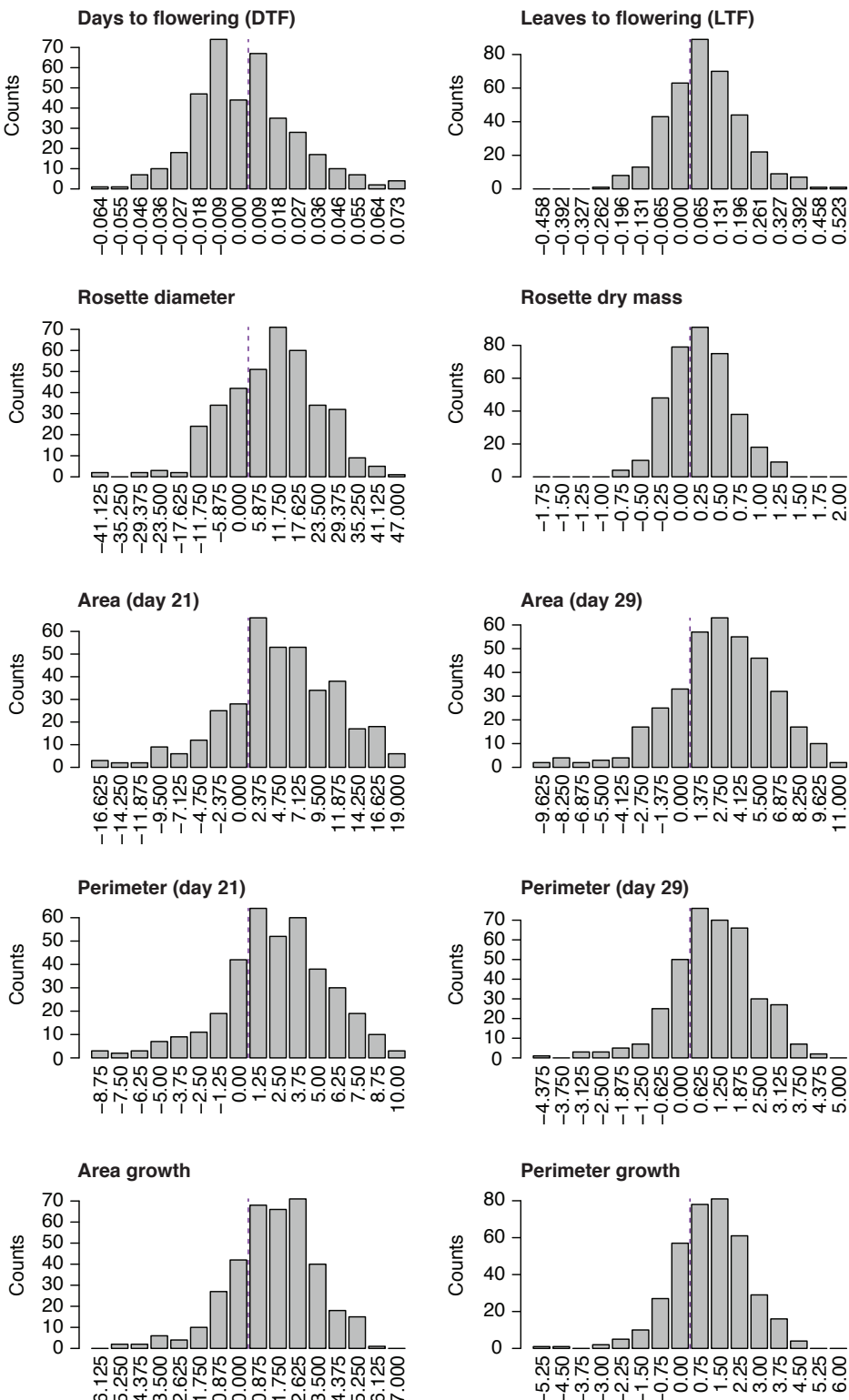
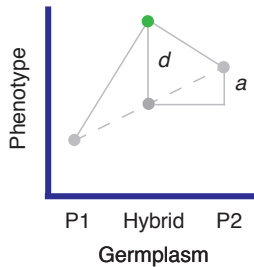
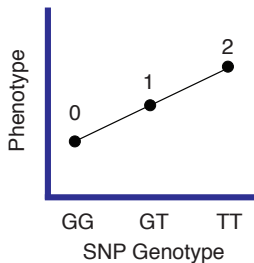
Figure S3

Figure S4

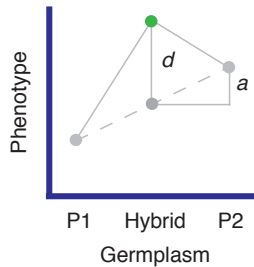
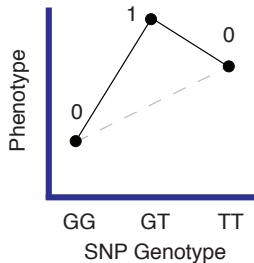
Experiment 1:

Model: Additive
Phenotype: Mean



Experiment 2:

Model: Overdominant
Phenotype: Mean



Experiment 3:

Model: Overdominant
Phenotype: d

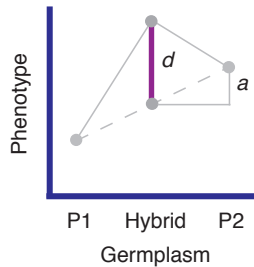
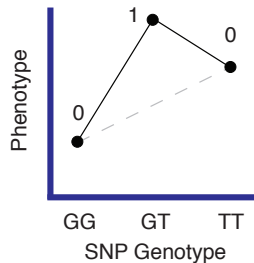


Figure S5

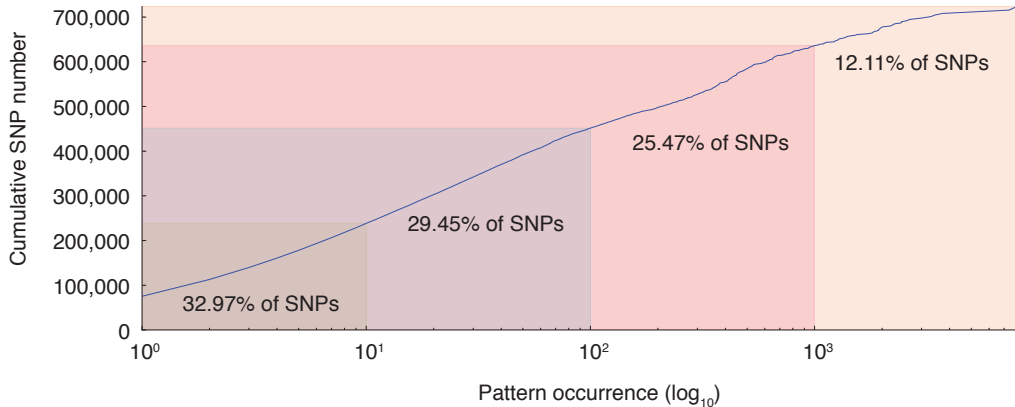


Figure S6

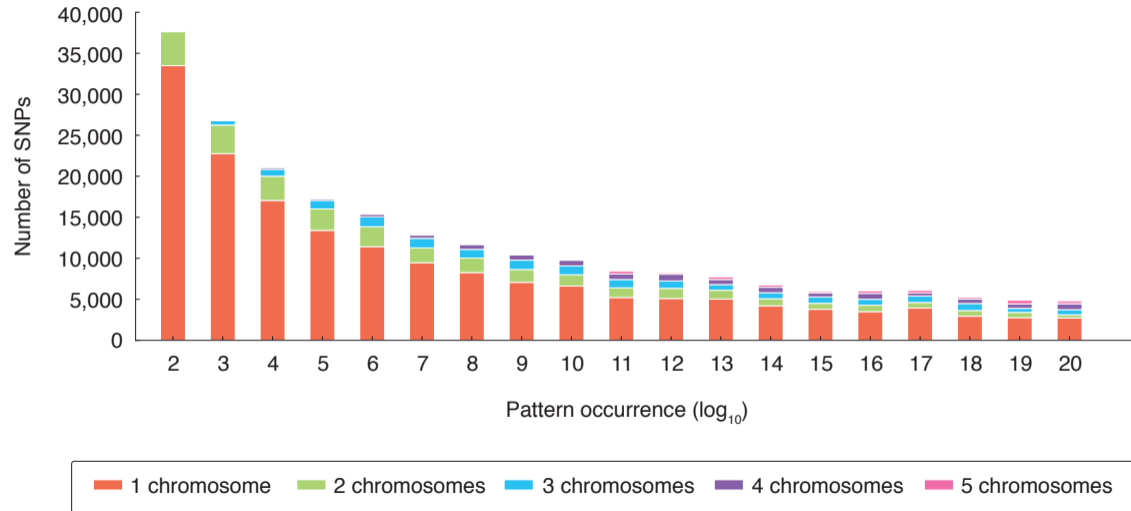


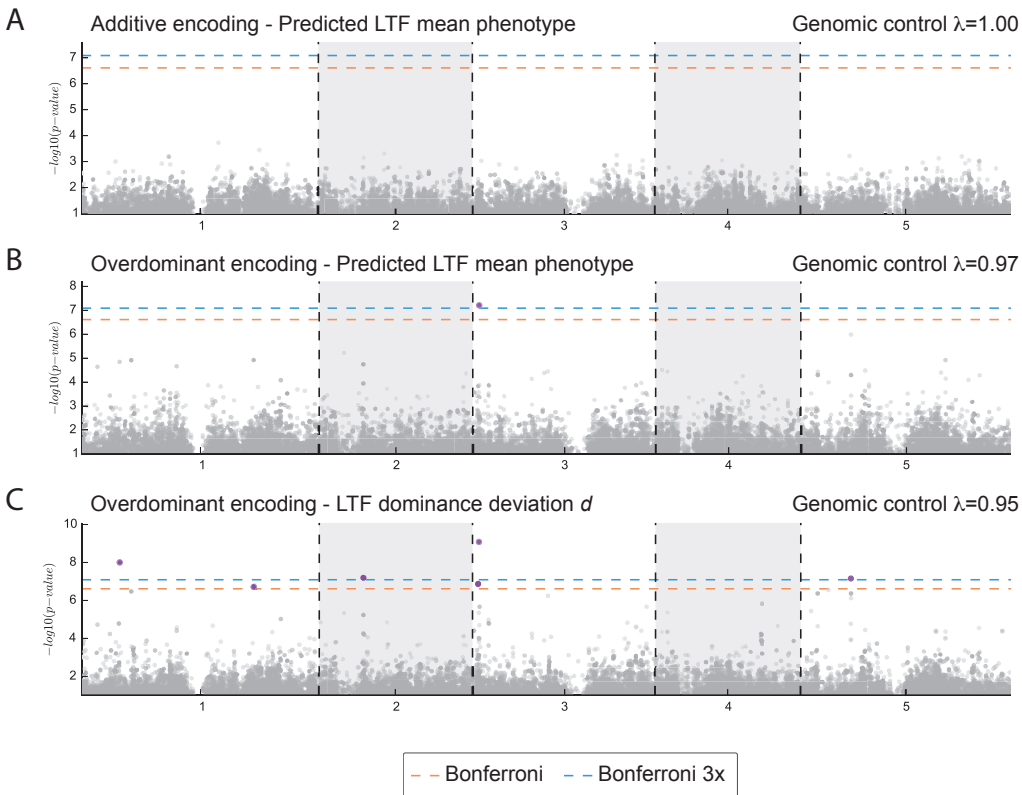
Figure S7

Figure S8

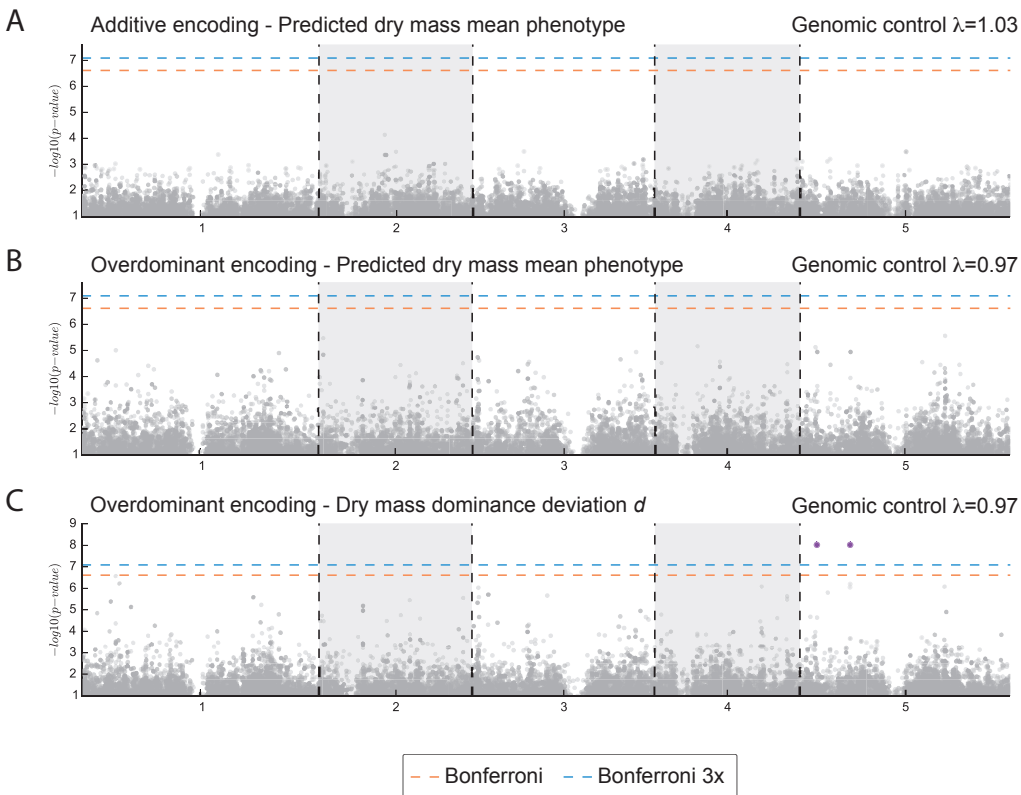


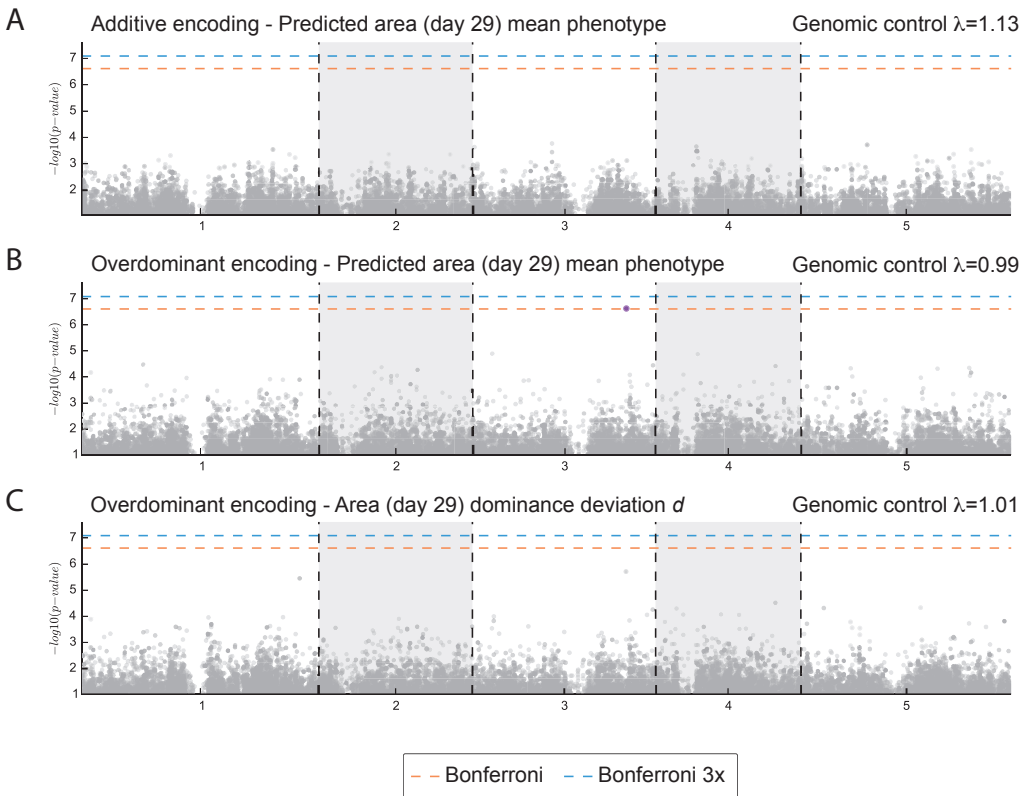
Figure S9

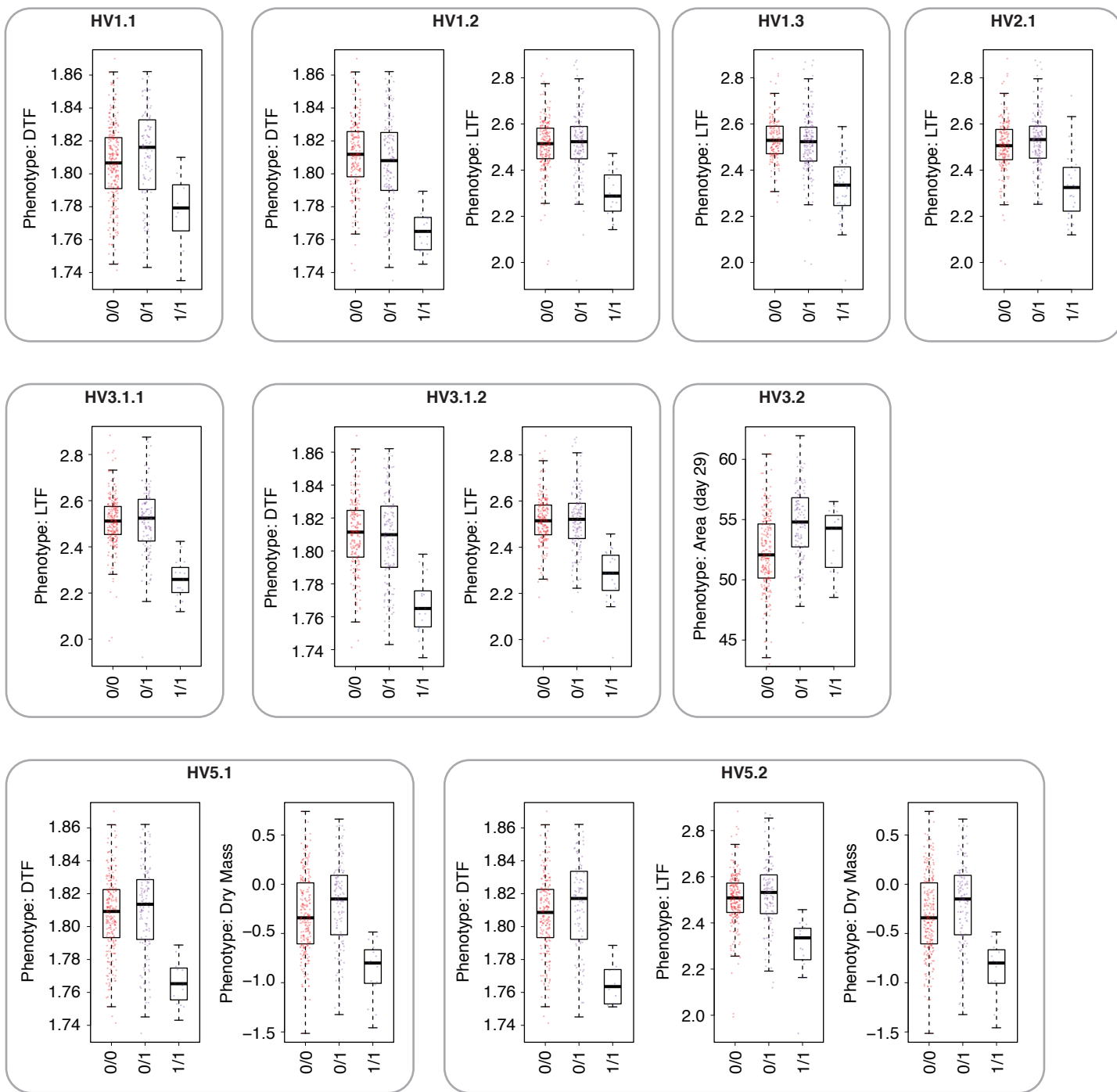
Figure S10

Figure S11

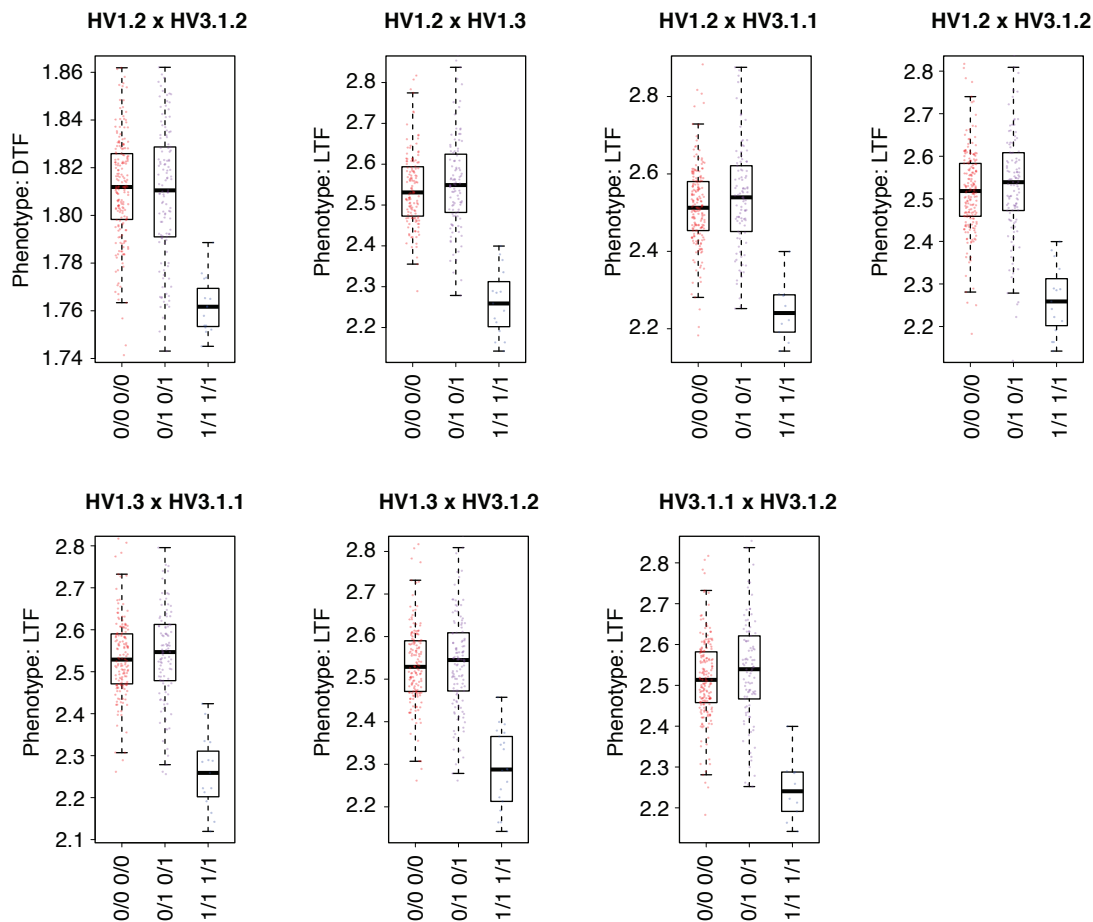


Figure S12

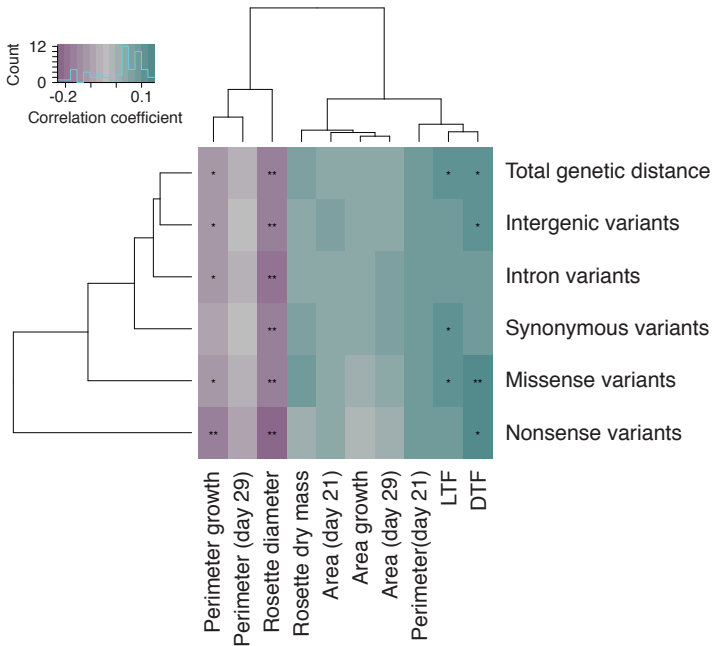
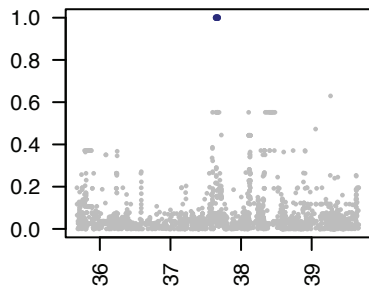
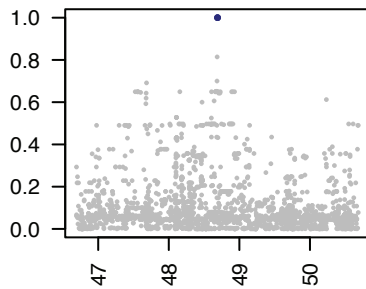


Figure S13

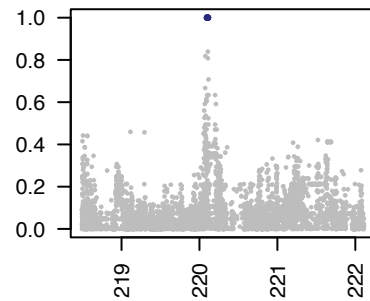
HV1.1



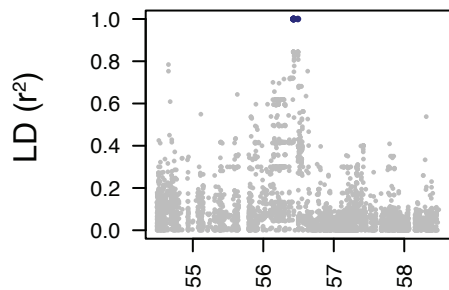
HV1.2



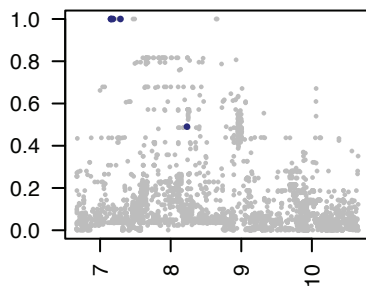
HV1.3



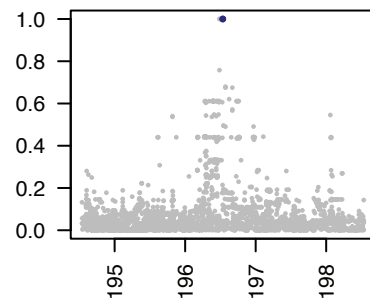
HV2.1



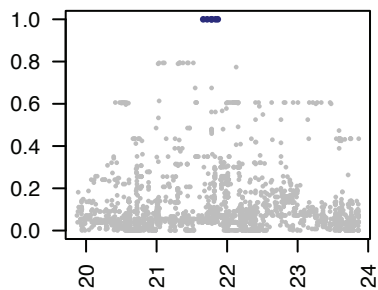
HV3.1.1/HV3.1.2



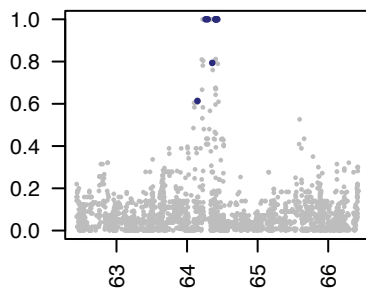
HV3.2



HV5.1



HV5.2



Chromosome Position (100 kb)

Figure S14

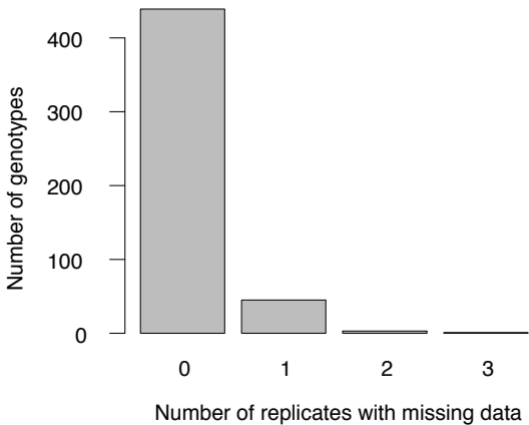


Figure S15

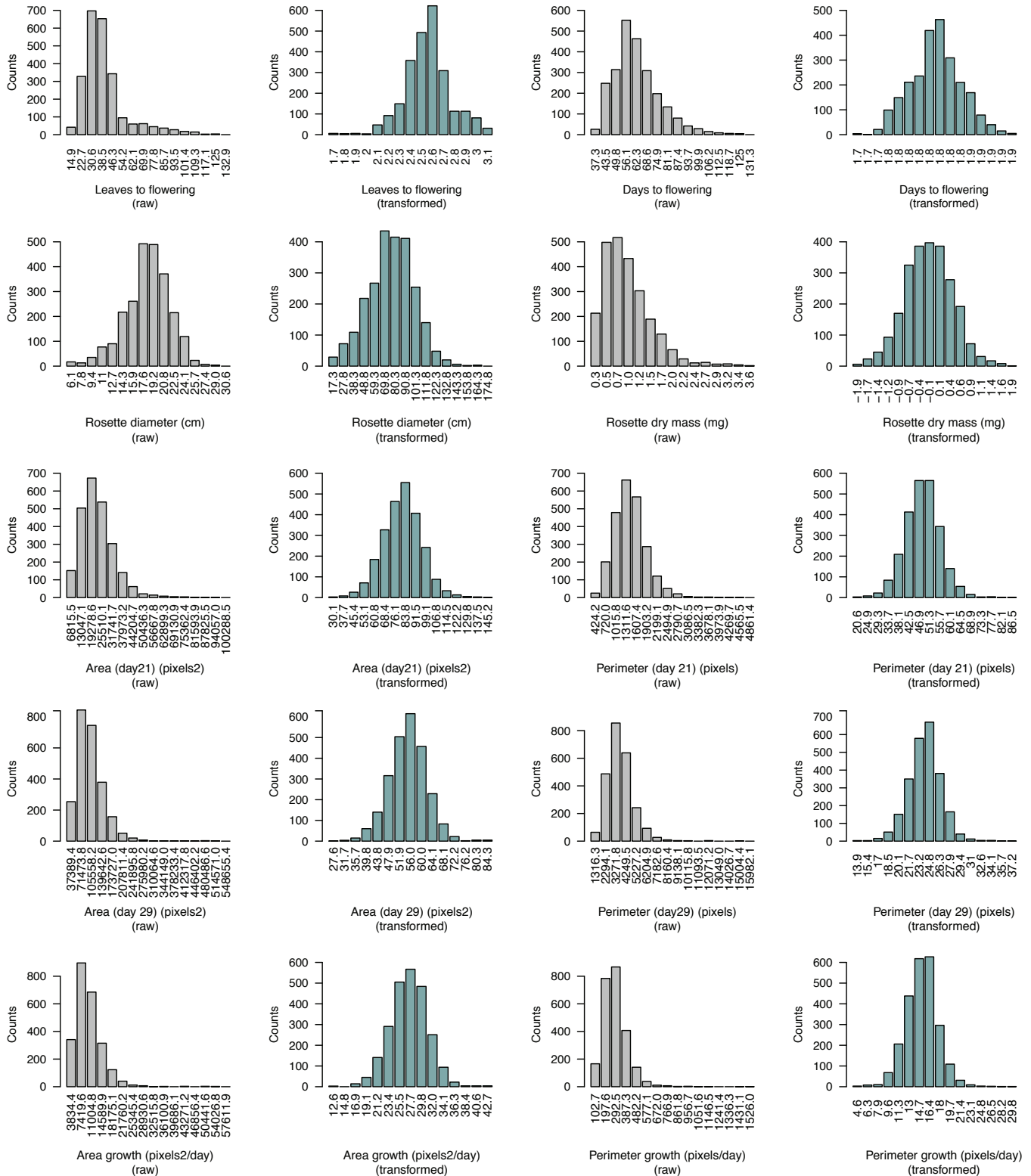


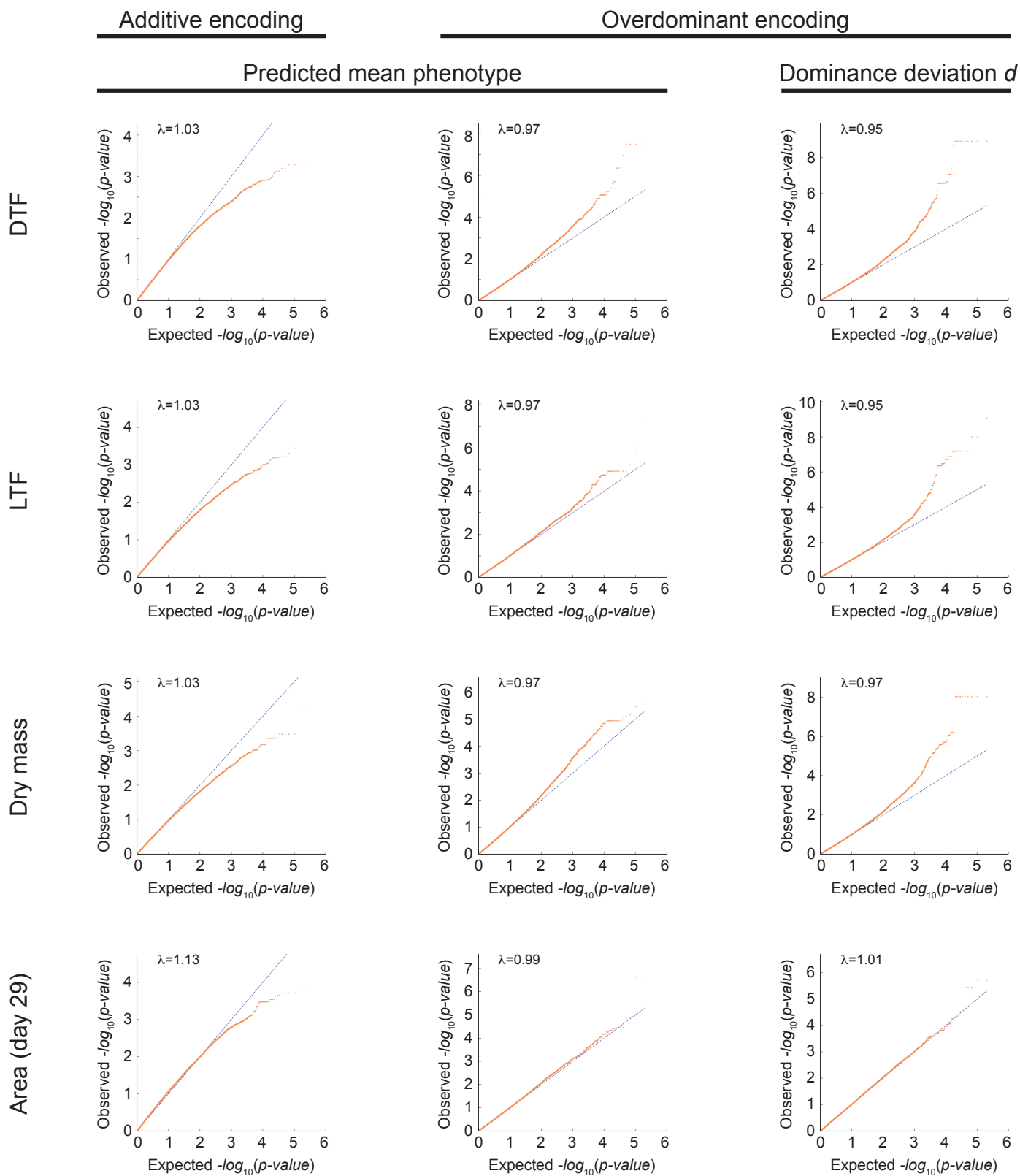
Figure S16

Table S1. Germplasm information.

Cross ID	Female parent	Male parent	Final data set (1=yes,0=no)
C001	ICE29	ICE63	1
C002	ICE29	ICE72	1
C003	ICE29	ICE79	1
C004	ICE29	ICE92	1
C005	ICE29	ICE107	1
C006	ICE29	ICE150	1
C007	ICE29	ICE212	1
C008	ICE29	Bak-2	1
C009	ICE29	Cdm-0	1
C010	ICE29	Ey15-2	1
C011	ICE29	HKT2.4	1
C012	ICE29	Mer-6	1
C013	ICE29	Nie1-2	1
C014	ICE29	Qui-0	1
C015	ICE29	Sha	1
C016	ICE29	Tuescha-9	1
C017	ICE29	TueWa1-2	1
C018	ICE50	ICE29	1
C019	ICE50	ICE61	1
C020	ICE50	ICE73	1
C021	ICE50	ICE79	1
C022	ICE50	ICE119	1
C023	ICE50	ICE150	1
C024	ICE50	ICE181	1
C025	ICE50	ICE212	1
C026	ICE50	ICE216	1
C027	ICE50	ICE228	1
C028	ICE50	Bak-2	1

C029	ICE50	Ey15-2	1
C030	ICE50	Fei-0	1
C031	ICE50	Mer-6	1
C032	ICE50	Nie1-2	1
C033	ICE50	Qui-0	1
C034	ICE50	Sha	1
C035	ICE50	Yeg-1	1
C036	ICE61	ICE29	1
C037	ICE61	ICE72	1
C038	ICE61	ICE92	1
C039	ICE61	ICE107	1
C040	ICE61	ICE119	1
C041	ICE61	ICE150	1
C042	ICE61	ICE228	0
C043	ICE61	Bak-2	1
C044	ICE61	Cdm-0	1
C045	ICE61	Ey15-2	1
C046	ICE61	Koch-1	1
C047	ICE61	Mer-6	1
C048	ICE61	Qui-0	1
C049	ICE61	Rue3-1-31	1
C050	ICE61	Tuescha-9	1
C051	ICE61	TueWa1-2	1
C052	ICE61	Yeg-1	1
C053	ICE63	ICE50	1
C054	ICE63	ICE61	1
C055	ICE63	ICE72	1
C056	ICE63	ICE73	1
C057	ICE63	ICE79	1
C058	ICE63	ICE107	1

C059	ICE63	ICE111	1
C060	ICE63	ICE212	1
C061	ICE63	Ey15-2	1
C062	ICE63	Fei-0	1
C063	ICE63	HKT2.4	1
C064	ICE63	Rue3-1-31	1
C065	ICE63	Sha	1
C066	ICE63	Tuescha-9	1
C067	ICE63	Yeg-1	1
C068	ICE72	ICE50	1
C069	ICE72	ICE73	1
C070	ICE72	ICE150	1
C071	ICE72	ICE212	1
C072	ICE72	ICE216	1
C073	ICE72	Bak-2	1
C074	ICE72	Cdm-0	1
C075	ICE72	Ey15-2	1
C076	ICE72	Fei-0	1
C077	ICE72	HKT2.4	1
C078	ICE72	Koch-1	1
C079	ICE72	Nie1-2	1
C080	ICE72	Qui-0	1
C081	ICE72	Tuescha-9	1
C082	ICE72	Yeg-1	1
C083	ICE73	ICE29	1
C084	ICE73	ICE61	1
C085	ICE73	ICE107	1
C086	ICE73	ICE111	1
C087	ICE73	ICE119	1
C088	ICE73	ICE181	1

C089	ICE73	ICE216	1
C090	ICE73	ICE228	1
C091	ICE73	Bak-2	1
C092	ICE73	Cdm-0	1
C093	ICE73	Ey15-2	1
C094	ICE73	HKT2.4	1
C095	ICE73	Koch-1	1
C096	ICE73	Mer-6	1
C097	ICE73	Nie1-2	1
C098	ICE73	Qui-0	1
C099	ICE73	Rue3-1-31	1
C100	ICE79	ICE61	1
C101	ICE79	ICE72	1
C102	ICE79	ICE73	1
C103	ICE79	ICE92	1
C104	ICE79	ICE107	1
C105	ICE79	ICE111	1
C106	ICE79	Bak-2	1
C107	ICE79	HKT2.4	1
C108	ICE79	Koch-1	1
C109	ICE79	Mer-6	1
C110	ICE79	Qui-0	1
C111	ICE79	Rue3-1-31	1
C112	ICE79	Tuescha-9	1
C113	ICE79	TueWa1-2	1
C114	ICE92	ICE50	1
C115	ICE92	ICE63	1
C116	ICE92	ICE72	1
C117	ICE92	ICE73	1
C118	ICE92	ICE107	1

C119	ICE92	ICE212	1
C120	ICE92	ICE216	1
C121	ICE92	ICE228	1
C122	ICE92	Bak-2	1
C123	ICE92	Cdm-0	1
C124	ICE92	HKT2.4	1
C125	ICE92	Mer-6	1
C126	ICE92	Qui-0	1
C127	ICE92	Sha	1
C128	ICE107	ICE50	1
C129	ICE107	ICE72	1
C130	ICE107	ICE111	1
C131	ICE107	ICE119	1
C132	ICE107	ICE150	1
C133	ICE107	ICE216	1
C134	ICE107	ICE228	1
C135	ICE107	Bak-2	1
C136	ICE107	Cdm-0	1
C137	ICE107	Ey15-2	1
C138	ICE107	HKT2.4	1
C139	ICE107	Qui-0	1
C140	ICE107	Rue3-1-31	1
C141	ICE107	Sha	1
C142	ICE107	Yeg-1	1
C143	ICE111	ICE29	1
C144	ICE111	ICE50	1
C145	ICE111	ICE61	1
C146	ICE111	ICE72	1
C147	ICE111	ICE92	1
C148	ICE111	ICE119	1

C149	ICE111	ICE150	1
C150	ICE111	Bak-2	1
C151	ICE111	Cdm-0	1
C152	ICE111	HKT2.4	1
C153	ICE111	Koch-1	1
C154	ICE111	Qui-0	1
C155	ICE119	ICE29	1
C156	ICE119	ICE63	1
C157	ICE119	ICE72	1
C158	ICE119	ICE79	1
C159	ICE119	ICE92	1
C160	ICE119	ICE150	1
C161	ICE119	Fei-0	1
C162	ICE119	HKT2.4	1
C163	ICE119	Koch-1	1
C164	ICE119	Mer-6	1
C165	ICE119	Nie1-2	1
C166	ICE119	Qui-0	1
C167	ICE119	Tuescha-9	1
C168	ICE119	TueWa1-2	1
C169	ICE119	Yeg-1	1
C170	ICE150	ICE63	1
C171	ICE150	ICE73	1
C172	ICE150	ICE79	1
C173	ICE150	ICE92	1
C174	ICE150	ICE181	1
C175	ICE150	ICE212	1
C176	ICE150	ICE228	1
C177	ICE150	Bak-2	1
C178	ICE150	Ey15-2	1

C179	ICE150	Koch-1	1
C180	ICE150	Nie1-2	1
C181	ICE150	Qui-0	1
C182	ICE150	Sha	1
C183	ICE150	Tuescha-9	1
C184	ICE181	ICE29	1
C185	ICE181	ICE61	1
C186	ICE181	ICE63	1
C187	ICE181	ICE72	1
C188	ICE181	ICE79	1
C189	ICE181	ICE92	1
C190	ICE181	ICE107	1
C191	ICE181	ICE111	1
C192	ICE181	ICE119	1
C193	ICE181	ICE216	1
C194	ICE181	Bak-2	1
C195	ICE181	Ey15-2	1
C196	ICE181	HKT2.4	1
C197	ICE181	Nie1-2	1
C198	ICE181	Rue3-1-31	1
C199	ICE212	ICE61	1
C200	ICE212	ICE73	1
C201	ICE212	ICE79	1
C202	ICE212	ICE107	1
C203	ICE212	ICE111	1
C204	ICE212	ICE119	1
C205	ICE212	ICE181	1
C206	ICE212	ICE216	1
C207	ICE212	Cdm-0	1
C208	ICE212	Ey15-2	1

C209	ICE212	Mer-6	1
C210	ICE212	Rue3-1-31	1
C211	ICE212	TueWa1-2	0
C212	ICE216	ICE29	1
C213	ICE216	ICE61	1
C214	ICE216	ICE63	1
C215	ICE216	ICE79	1
C216	ICE216	ICE111	1
C217	ICE216	ICE119	1
C218	ICE216	ICE150	1
C219	ICE216	ICE228	1
C220	ICE216	Bak-2	1
C221	ICE216	Ey15-2	1
C222	ICE216	HKT2.4	1
C223	ICE216	Nie1-2	1
C224	ICE216	Qui-0	1
C225	ICE216	Rue3-1-31	1
C226	ICE216	Yeg-1	1
C227	ICE228	ICE29	1
C228	ICE228	ICE63	1
C229	ICE228	ICE72	1
C230	ICE228	ICE79	1
C231	ICE228	ICE111	1
C232	ICE228	ICE119	1
C233	ICE228	ICE181	1
C234	ICE228	ICE212	1
C235	ICE228	Bak-2	1
C236	ICE228	Cdm-0	1
C237	ICE228	Koch-1	1
C238	ICE228	Rue3-1-31	1

C239	ICE228	Sha	1
C240	ICE228	TueWa1-2	1
C241	ICE228	Yeg-1	1
C242	Bak-2	ICE63	1
C243	Bak-2	ICE119	1
C244	Bak-2	ICE212	1
C245	Bak-2	Cdm-0	1
C246	Bak-2	Ey15-2	1
C247	Bak-2	Fei-0	1
C248	Bak-2	Mer-6	1
C249	Bak-2	Nie1-2	1
C250	Bak-2	Qui-0	1
C251	Bak-2	TueWa1-2	1
C252	Bak-2	Yeg-1	1
C253	Cdm-0	ICE50	1
C254	Cdm-0	ICE63	1
C255	Cdm-0	ICE79	1
C256	Cdm-0	ICE119	1
C257	Cdm-0	ICE150	1
C258	Cdm-0	ICE181	1
C259	Cdm-0	ICE216	1
C260	Cdm-0	Ey15-2	1
C261	Cdm-0	Koch-1	1
C262	Cdm-0	Mer-6	1
C263	Cdm-0	Rue3-1-31	1
C264	Cdm-0	Sha	1
C265	Cdm-0	TueWa1-2	1
C266	Cdm-0	Yeg-1	0
C267	Ey15-2	ICE79	1
C268	Ey15-2	ICE92	1

C269	Ey15-2	ICE111	1
C270	Ey15-2	ICE119	1
C271	Ey15-2	ICE228	1
C272	Ey15-2	Fei-0	1
C273	Ey15-2	Mer-6	1
C274	Ey15-2	Nie1-2	1
C275	Ey15-2	Qui-0	1
C276	Ey15-2	Rue3-1-31	1
C277	Ey15-2	Tuescha-9	1
C278	Ey15-2	TueWa1-2	1
C279	Ey15-2	Yeg-1	1
C280	Fei-0	ICE29	1
C281	Fei-0	ICE61	1
C282	Fei-0	ICE73	1
C283	Fei-0	ICE79	1
C284	Fei-0	ICE92	1
C285	Fei-0	ICE107	1
C286	Fei-0	ICE111	1
C287	Fei-0	ICE150	1
C288	Fei-0	ICE181	1
C289	Fei-0	ICE212	1
C290	Fei-0	ICE216	1
C291	Fei-0	ICE228	1
C292	Fei-0	Cdm-0	1
C293	Fei-0	TueWa1-2	1
C294	HKT2.4	ICE50	1
C295	HKT2.4	ICE61	1
C296	HKT2.4	ICE150	1
C297	HKT2.4	ICE212	1
C298	HKT2.4	ICE228	1

C299	HKT2.4	Bak-2	1
C300	HKT2.4	Cdm-0	1
C301	HKT2.4	Ey15-2	1
C302	HKT2.4	Fei-0	1
C303	HKT2.4	Koch-1	1
C304	HKT2.4	Rue3-1-31	1
C305	HKT2.4	TueWa1-2	1
C306	HKT2.4	Yeg-1	1
C307	Koch-1	ICE29	1
C308	Koch-1	ICE50	1
C309	Koch-1	ICE63	1
C310	Koch-1	ICE92	1
C311	Koch-1	ICE107	1
C312	Koch-1	ICE181	1
C313	Koch-1	ICE212	1
C314	Koch-1	ICE216	1
C315	Koch-1	Bak-2	1
C316	Koch-1	Ey15-2	1
C317	Koch-1	Fei-0	1
C318	Koch-1	Mer-6	1
C319	Koch-1	Sha	1
C320	Koch-1	Yeg-1	1
C321	Mer-6	ICE63	1
C322	Mer-6	ICE72	1
C323	Mer-6	ICE107	1
C324	Mer-6	ICE111	1
C325	Mer-6	ICE150	0
C326	Mer-6	ICE181	1
C327	Mer-6	ICE216	1
C328	Mer-6	ICE228	1

C329	Mer-6	Fei-0	1
C330	Mer-6	HKT2.4	1
C331	Mer-6	Qui-0	1
C332	Mer-6	Rue3-1-31	1
C333	Mer-6	Sha	1
C334	Nie1-2	ICE61	1
C335	Nie1-2	ICE63	1
C336	Nie1-2	ICE79	1
C337	Nie1-2	ICE92	1
C338	Nie1-2	ICE107	1
C339	Nie1-2	ICE111	1
C340	Nie1-2	ICE212	1
C341	Nie1-2	ICE228	1
C342	Nie1-2	Cdm-0	1
C343	Nie1-2	Fei-0	1
C344	Nie1-2	HKT2.4	1
C345	Nie1-2	Koch-1	1
C346	Nie1-2	Mer-6	1
C347	Nie1-2	Qui-0	1
C348	Nie1-2	Rue3-1-31	1
C349	Qui-0	ICE63	1
C350	Qui-0	ICE181	1
C351	Qui-0	ICE212	1
C352	Qui-0	ICE228	1
C353	Qui-0	Cdm-0	1
C354	Qui-0	Fei-0	1
C355	Qui-0	HKT2.4	1
C356	Qui-0	Koch-1	1
C357	Qui-0	Rue3-1-31	1
C358	Qui-0	Sha	1

C359	Qui-0	Tuescha-9	1
C360	Qui-0	TueWa1-2	1
C361	Qui-0	Yeg-1	1
C362	Rue3-1-31	ICE29	1
C363	Rue3-1-31	ICE50	1
C364	Rue3-1-31	ICE72	1
C365	Rue3-1-31	ICE92	1
C366	Rue3-1-31	ICE111	1
C367	Rue3-1-31	ICE119	1
C368	Rue3-1-31	ICE150	1
C369	Rue3-1-31	Bak-2	1
C370	Rue3-1-31	Fei-0	1
C371	Rue3-1-31	Koch-1	1
C372	Rue3-1-31	Sha	1
C373	Rue3-1-31	Tuescha-9	1
C374	Rue3-1-31	TueWa1-2	1
C375	Sha	ICE61	1
C376	Sha	ICE72	1
C377	Sha	ICE73	1
C378	Sha	ICE79	1
C379	Sha	ICE111	1
C380	Sha	ICE119	1
C381	Sha	ICE181	1
C382	Sha	ICE212	1
C383	Sha	ICE216	1
C384	Sha	Bak-2	1
C385	Sha	Ey15-2	1
C386	Sha	Fei-0	1
C387	Sha	HKT2.4	1
C388	Sha	Nie1-2	1

C389	Sha	TueWa1-2	0
C390	Tuescha-9	ICE50	1
C391	Tuescha-9	ICE73	1
C392	Tuescha-9	ICE92	1
C393	Tuescha-9	ICE107	1
C394	Tuescha-9	ICE111	1
C395	Tuescha-9	ICE181	1
C396	Tuescha-9	ICE212	1
C397	Tuescha-9	ICE216	1
C398	Tuescha-9	ICE228	1
C399	Tuescha-9	Bak-2	1
C400	Tuescha-9	Cdm-0	0
C401	Tuescha-9	Fei-0	1
C402	Tuescha-9	HKT2.4	1
C403	Tuescha-9	Koch-1	1
C404	Tuescha-9	Mer-6	1
C405	Tuescha-9	Nie1-2	1
C406	Tuescha-9	Sha	1
C407	TueWa1-2	ICE50	1
C408	TueWa1-2	ICE63	1
C409	TueWa1-2	ICE72	1
C410	TueWa1-2	ICE73	1
C411	TueWa1-2	ICE92	1
C412	TueWa1-2	ICE107	1
C413	TueWa1-2	ICE111	1
C414	TueWa1-2	ICE150	1
C415	TueWa1-2	ICE181	1
C416	TueWa1-2	ICE216	0
C417	TueWa1-2	Koch-1	1
C418	TueWa1-2	Mer-6	1

C419	TueWa1-2	Nie1-2	1
C420	TueWa1-2	Tuescha-9	1
C421	TueWa1-2	Yeg-1	1
C422	Yeg-1	ICE29	1
C423	Yeg-1	ICE73	1
C424	Yeg-1	ICE79	1
C425	Yeg-1	ICE92	1
C426	Yeg-1	ICE111	1
C427	Yeg-1	ICE150	1
C428	Yeg-1	ICE181	1
C429	Yeg-1	ICE212	1
C430	Yeg-1	Fei-0	1
C431	Yeg-1	Mer-6	1
C432	Yeg-1	Nie1-2	1
C433	Yeg-1	Rue3-1-31	1
C434	Yeg-1	Sha	1
C435	Yeg-1	Tuescha-9	1
PM001	ICE29	ICE29	1
PM002	ICE50	ICE50	1
PM003	ICE61	ICE61	0
PM004	ICE63	ICE63	1
PM005	ICE72	ICE72	1
PM006	ICE73	ICE73	1
PM007	ICE79	ICE79	1
PM008	ICE92	ICE92	1
PM009	ICE107	ICE107	1
PM010	ICE111	ICE111	1
PM011	ICE119	ICE119	1
PM012	ICE150	ICE150	1
PM013	ICE181	ICE181	1

PM014	ICE212	ICE212	1
PM015	ICE216	ICE216	1
PM016	ICE228	ICE228	1
PM017	Bak-2	Bak-2	0
PM018	Cdm-0	Cdm-0	1
PM019	Ey15-2	Ey15-2	1
PM020	Fei-0	Fei-0	1
PM021	HKT2.4	HKT2.4	1
PM022	Koch-1	Koch-1	1
PM023	Mer-6	Mer-6	1
PM024	Nie1-2	Nie1-2	1
PM025	Qui-0	Qui-0	1
PM026	Rue3-1-31	Rue3-1-31	1
PM027	Sha	Sha	1
PM028	Tuescha-9	Tuescha-9	1
PM029	TueWa1-2	TueWa1-2	1
PM030	Yeg-1	Yeg-1	1
PS001	ICE29	ICE29	1
PS002	ICE50	ICE50	1
PS003	ICE61	ICE61	1
PS004	ICE63	ICE63	1
PS005	ICE72	ICE72	1
PS006	ICE73	ICE73	1
PS007	ICE79	ICE79	1
PS008	ICE92	ICE92	1
PS009	ICE107	ICE107	1
PS010	ICE111	ICE111	1
PS011	ICE119	ICE119	1
PS012	ICE150	ICE150	1
PS013	ICE181	ICE181	1

PS014	ICE212	ICE212	1
PS015	ICE216	ICE216	1
PS016	ICE228	ICE228	1
PS017	Bak-2	Bak-2	1
PS018	Cdm-0	Cdm-0	1
PS019	Ey15-2	Ey15-2	1
PS020	Fei-0	Fei-0	1
PS021	HKT2.4	HKT2.4	1
PS022	Koch-1	Koch-1	1
PS023	Mer-6	Mer-6	1
PS024	Nie1-2	Nie1-2	1
PS025	Qui-0	Qui-0	1
PS026	Rue3-1-31	Rue3-1-31	1
PS027	Sha	Sha	1
PS028	Tuescha-9	Tuescha-9	1
PS029	TueWa1-2	TueWa1-2	1
PS030	Yeg-1	Yeg-1	1

Table S2. GCA, SCA, and heritability estimates.

Phenotype	Linear mixed model estimates (SAS)							Linear model estimates (R)	
	GCA	SCA	Residual variance	Total phenotypic variance	Narrow-sense heritability (h ²)	Standard error (h ²)	Broad-sense heritability (H ²)	Standard error (H ²)	Broad-sense heritability (H ²)
DTF	0.00026	0.00033	0.00023	0.00107	0.47916	0.07013	0.78346	0.02971	0.78095
LTF	0.00690	0.01168	0.01534	0.04082	0.33824	0.06422	0.62433	0.03813	0.63450
Rosette diameter	95.62288	122.33440	195.19542	508.77558	0.37589	0.06620	0.61634	0.04198	0.62977
Rosette dry mass	0.05973	0.10434	0.10238	0.32618	0.36623	0.06657	0.68612	0.03437	0.68275
Area (day 21)	7.65954	38.61871	146.83454	200.77233	0.07630	0.02471	0.26865	0.02733	0.30500
Perimeter (day 21)	1.58477	10.39032	40.86399	54.42385	0.05824	0.02074	0.24915	0.02561	0.28090
Area (day 29)	3.61674	8.60298	30.35852	46.19497	0.15659	0.04014	0.34282	0.03551	0.35842
Perimeter (day 29)	0.56633	1.13805	3.59984	5.87055	0.19294	0.04619	0.38680	0.03853	0.39222
Area growth	1.25570	2.35933	7.92207	12.79280	0.19631	0.04659	0.38074	0.03918	0.38833
Perimeter growth	0.96668	1.27079	3.49608	6.70024	0.28855	0.05850	0.47822	0.04488	0.47010

Table S3. Summary of significant SNPs detected in GWASs.

	<u>Additive encoding</u>		<u>Overdominant encoding</u>		<u>Additive encoding</u>		<u>Overdominant encoding</u>	
	<u>Predicted mean phenotype</u>		<u>Dominance deviation d</u>		<u>Predicted mean phenotype</u>		<u>Dominance deviation d</u>	
	<u>Within experiment Bonferroni correction</u>				<u>Across experiment (3) Bonferroni correction</u>			
DTF	0	5	19		0	4	13	
LTF	0	1	21		0	1	12	
Dry mass	0	0	10		0	0	10	
Area (day 29)	0	2	0		0	0	0	

Table S4. Summary of significant SNPs per phenotype.

SNP ID	Region ID	<u>Additive encoding</u>				<u>Overdominant encoding</u>				<u>Additive encoding</u>				<u>Overdominant encoding</u>											
		<u>Predicted mean phenotype</u>				<u>Dominance deviation d</u>				<u>Predicted mean phenotype</u>				<u>Dominance deviation d</u>											
		<u>Within experiment Bonferroni correction</u>								<u>Across experiment (3) Bonferroni correction</u>															
		<u>DTE</u>	<u>LTE</u>	<u>Dry mass</u>	<u>Area (day 29)</u>	<u>DTE</u>	<u>LTE</u>	<u>Dry mass</u>	<u>Area (day 29)</u>	<u>DTE</u>	<u>LTE</u>	<u>Dry mass</u>	<u>Area (day 29)</u>	<u>DTE</u>	<u>LTE</u>	<u>Dry mass</u>	<u>Area (day 29)</u>	<u>DTE</u>	<u>LTE</u>	<u>Dry mass</u>	<u>Area (day 29)</u>	<u>DTE</u>	<u>LTE</u>	<u>Dry mass</u>	<u>Area (day 29)</u>
1_3764724	1	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
1_3764778	1	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
1_3766766	1	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
1_3766808	1	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
1_4869029	2	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
1_4869068	2	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
1_22009862	3	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
1_22009873	3	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
1_22009978	3	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
1_22010004	3	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
1_22010908	3	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
2_5642726	4	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
2_5642735	4	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
2_5642738	4	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
2_5642877	4	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
2_5642901	4	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
2_5643184	4	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
2_5643193	4	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
2_5649084	4	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
3_715448	5	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
3_715567	5	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
3_718041	5	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
3_728908	5	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
3_823046	6	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE

Table S5. Summary of SConES SNPs.

Phenotype	SNP encoding	Phenotypic component	Number of SNPs
DTF	Overdominant	Fitted mean	2
DTF	Overdominant	Dominance Deviation <i>d</i>	72
LTF	Overdominant	Fitted mean	66
LTF	Overdominant	Dominance Deviation <i>d</i>	88
Dry mass	Overdominant	Fitted mean	5
Dry mass	Overdominant	Dominance Deviation <i>d</i>	190
Rosette diameter	Overdominant	Fitted mean	111
Rosette diameter	Overdominant	Dominance Deviation <i>d</i>	19
Area (day 21)	Overdominant	Fitted mean	39
Area (day 21)	Overdominant	Dominance Deviation <i>d</i>	121
Area (day 29)	Overdominant	Fitted mean	0
Area (day 29)	Overdominant	Dominance Deviation <i>d</i>	10
Perimeter (day 21)	Overdominant	Fitted mean	2
Perimeter (day 21)	Overdominant	Dominance Deviation <i>d</i>	30
Perimeter (day 29)	Overdominant	Fitted mean	13
Perimeter (day 29)	Overdominant	Dominance Deviation <i>d</i>	13
Area growth	Overdominant	Fitted mean	0
Area growth	Overdominant	Dominance Deviation <i>d</i>	64
Perimeter growth	Overdominant	Fitted mean	9
Perimeter growth	Overdominant	Dominance Deviation <i>d</i>	324

Table S7. Summary of candidate regions.

<u>Region Number</u>	<u>Locus ID</u>	<u>Chr</u>	<u>Start</u>	<u>Stop</u>	<u>LD block region</u>	<u>Reference SNP</u>	<u>Associated phenotypes</u>
1	HV1.1	1	3764724	3766808	Chr1:3764724..3766808	3768662	DTF mean, DTF <i>d</i>
2	HV1.2	1	4869029	4869068	Chr1:4868390..4869079	4869079	DTF <i>d</i> , LTF <i>d</i>
3	HV1.3	1	22009862	22010908	Chr1:22008152..22010924	22010924	LTF <i>d</i>
4	HV2.1	2	5642726	5649084	Chr2:5642726..5649084	5649084	LTF <i>d</i>
5	HV3.1.1	3	715448	728908	Chr3:715448..865775	756657	LTF <i>d</i>
6	HV3.1.2	3	823046	823046	Chr3:715448..865775	756657	DTF <i>d</i> , LTF mean, LTF <i>d</i>
7	HV3.2	3	19653355	19653356	Chr3:19648666..19653356	19653356	Area (day 29) mean
8	HV5.1	5	2166140	2183630	Chr5:2166140..2186738	2186738	DTF <i>d</i> , Dry mass <i>d</i>
9	HV5.2	5	6414746	6441707	Chr5:6414746..6442843	6442843	DTF mean, DTF <i>d</i> , LTF <i>d</i> , Dry mass <i>d</i>

<u>Phenotypic behavior with trait mean</u>	<u>Candidate</u>	<u>Annotation</u>	<u>Type</u>
Overdominant	AT1G11230	Unknown Protein	Promoter
Dominant	AT1G14250	GDA1/CD39 nucleoside phosphatase	Gene
Dominant	AT1G59810	AGL50	Gene
Tending towards overdominant	NA	NA	NA
Dominant	NA	NA	NA
Dominant	NA	NA	NA
Tending towards overdominant	AT3G52990	Pyruvate kinase family protein	Gene
Overdominant	NA	NA	NA
Overdominant	NA	NA	NA

Table S8. Gene Ontology (GO) enrichment.

	<u>Predicted mean phenotype</u>									
	<u>DTE</u>	<u>LTE</u>	<u>Rosette dry mass</u>	<u>Rosette diameter</u>	<u>Area (day 21)</u>	<u>Area (day 29)</u>	<u>Perimeter (day 21)</u>	<u>Perimeter (day 29)</u>	<u>Area growth</u>	<u>Perimeter growth</u>
carbohydrate metabolic process	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.04E-07	2.76E-05	1.33E-08	1.00E+00	1.00E+00	1.00E+00
starch metabolic process	1.00E+00	1.00E+00	1.00E+00	1.00E+00	2.85E-06	5.83E-05	1.38E-08	1.00E+00	1.00E+00	1.00E+00
cellular response to water deprivation	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.30E-05	1.00E+00	3.10E-09	1.00E+00	1.00E+00	1.00E+00
polysaccharide metabolic process	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.87E-05	7.51E-03	1.35E-05	1.00E+00	1.00E+00	1.00E+00
mitochondrial ATP synthesis coupled electron transport	1.00E+00	1.00E+00	1.00E+00	1.00E+00	4.03E-05	4.84E-05	2.78E-05	1.15E-04	2.69E-05	6.59E-05
response to water deprivation	1.00E+00	1.00E+00	1.00E+00	1.00E+00	5.90E-05	1.00E+00	5.02E-09	1.00E+00	1.00E+00	1.00E+00
starch biosynthetic process	1.00E+00	1.00E+00	1.00E+00	1.00E+00	6.29E-05	1.00E+00	2.80E-07	1.00E+00	1.00E+00	1.00E+00
response to water	1.00E+00	1.00E+00	1.00E+00	1.00E+00	8.33E-05	1.00E+00	9.40E-09	1.00E+00	1.00E+00	1.00E+00
oligosaccharide metabolic process	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.52E-04	1.62E-05	4.21E-06	5.39E-03	2.14E-02	1.00E+00
cellular glucan metabolic process	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.53E-04	8.36E-05	1.25E-06	1.77E-02	2.25E-04	1.23E-04
glucan biosynthetic process	1.00E+00	1.00E+00	1.00E+00	1.00E+00	2.19E-04	2.81E-03	1.00E-06	1.00E+00	6.82E-03	4.53E-03
ATP synthesis coupled electron transport	1.00E+00	1.00E+00	1.00E+00	1.00E+00	3.84E-04	1.34E-04	3.12E-04	4.27E-04	1.70E-04	1.74E-04
pectin catabolic process	1.07E-02	1.00E+00	1.00E+00	1.00E+00	4.46E-04	1.00E+00	2.59E-04	1.00E+00	1.00E+00	1.00E+00
oxidative phosphorylation	1.00E+00	1.00E+00	1.00E+00	1.00E+00	6.05E-04	8.29E-04	4.45E-04	6.86E-04	6.83E-04	5.23E-04
stomatal complex morphogenesis	1.00E+00	1.00E+00	1.00E+00	1.00E+00	6.41E-04	1.00E+00	3.35E-02	1.00E+00	1.00E+00	1.00E+00
maltose metabolic process	1.00E+00	1.00E+00	1.00E+00	1.00E+00	7.52E-04	2.65E-02	5.06E-06	1.00E+00	1.00E+00	1.00E+00
pentose-phosphate shunt	1.00E+00	1.00E+00	1.00E+00	1.00E+00	8.26E-04	8.68E-04	2.48E-05	1.00E+00	1.00E+00	1.00E+00
myo-inositol hexakisphosphate biosynthetic process	1.00E+00	1.00E+00	1.00E+00	1.00E+00	8.26E-04	1.00E+00	5.98E-04	1.00E+00	1.00E+00	1.00E+00
apoptotic process	1.00E+00	1.00E+00	1.00E+00	1.46E-05	1.15E-03	1.00E+00	7.72E-04	1.00E+00	1.00E+00	1.00E+00
polysaccharide catabolic process	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.58E-03	1.00E+00	2.00E-03	1.00E+00	1.00E+00	1.00E+00
tyrosine metabolic process	1.00E+00	1.00E+00	1.00E+00	1.00E+00	2.03E-03	1.00E+00	1.34E-03	1.00E+00	1.00E+00	1.00E+00
shoot system morphogenesis	1.00E+00	1.00E+00	1.00E+00	1.00E+00	2.03E-03	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
respiratory electron transport chain	1.00E+00	1.00E+00	1.00E+00	1.00E+00	2.60E-03	1.42E-03	1.69E-03	1.84E-03	1.75E-03	1.86E-03
syncytium formation	1.00E+00	1.00E+00	1.00E+00	1.00E+00	3.35E-03	9.03E-03	2.18E-03	4.16E-02	4.46E-02	3.91E-02
glucose 6-phosphate metabolic process	1.00E+00	1.00E+00	1.00E+00	1.00E+00	4.26E-03	6.10E-04	1.67E-04	1.00E+00	1.00E+00	1.00E+00
NADP metabolic process	1.00E+00	1.00E+00	1.00E+00	1.00E+00	4.35E-03	1.37E-03	1.86E-04	1.00E+00	1.00E+00	1.00E+00

chromosome condensation	1.00E+00	1.00E+00	1.00E+00	1.00E+00	6.02E-03	1.00E+00	4.08E-03	1.00E+00	1.00E+00	1.00E+00
cellular carbohydrate metabolic process	1.00E+00	1.00E+00	1.00E+00	1.00E+00	6.72E-03	2.90E-02	1.14E-03	1.00E+00	3.71E-02	1.00E+00
cell-cell signaling involved in cell fate commitment	1.00E+00	1.00E+00	1.00E+00	1.00E+00	7.12E-03	1.00E+00	4.57E-03	1.00E+00	1.00E+00	1.00E+00
pectin metabolic process	6.30E-03	1.00E+00	1.00E+00	1.00E+00	8.19E-03	1.00E+00	4.08E-03	1.00E+00	1.00E+00	1.00E+00
regulation of meristem development	1.00E+00	1.00E+00	1.00E+00	1.00E+00	8.54E-03	1.00E+00	2.27E-02	1.00E+00	4.79E-02	1.00E+00
tricarboxylic acid cycle	1.00E+00	1.00E+00	1.00E+00	1.00E+00	8.54E-03	1.00E+00	5.33E-03	1.00E+00	1.00E+00	1.03E-02
polyamine catabolic process	1.90E-02	1.00E+00	2.27E-05	1.00E+00	8.61E-03	2.37E-02	1.00E+00	1.00E+00	1.00E+00	1.00E+00
cellular aldehyde metabolic process	1.00E+00	1.00E+00	1.00E+00	1.00E+00	8.72E-03	1.00E+00	2.86E-03	1.00E+00	1.00E+00	1.00E+00
ATP metabolic process	1.00E+00	1.00E+00	1.00E+00	1.00E+00	8.82E-03	1.00E+00	6.77E-03	3.23E-06	6.66E-06	3.72E-06
citrate metabolic process	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.01E-02	1.00E+00	7.98E-03	1.00E+00	1.00E+00	1.05E-02
response to abiotic stimulus	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.28E-02	1.00E+00	3.59E-06	1.00E+00	1.00E+00	1.00E+00
regulation of stomatal movement	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.51E-02	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
ammonium transport	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.66E-02	1.00E+00	1.37E-02	1.00E+00	1.00E+00	1.00E+00
basic amino acid transport	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.75E-02	1.00E+00	1.41E-02	1.00E+00	1.00E+00	1.00E+00
phospholipid transport	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.76E-02	1.00E+00	2.67E-03	1.00E+00	1.00E+00	1.00E+00
stomatal complex development	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.78E-02	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
protein homooligomerization	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.90E-02	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
regulation of meristem growth	1.00E+00	1.00E+00	1.00E+00	1.00E+00	2.16E-02	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
threonine catabolic process	1.00E+00	1.00E+00	1.00E+00	1.00E+00	2.45E-02	1.00E+00	2.10E-02	1.00E+00	1.00E+00	1.00E+00
cellular metal ion homeostasis	1.00E+00	1.00E+00	1.00E+00	1.00E+00	2.45E-02	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
aromatic amino acid family biosynthetic process, prephenate pathway	1.00E+00	1.00E+00	1.00E+00	1.00E+00	2.45E-02	1.00E+00	2.10E-02	1.00E+00	1.00E+00	1.00E+00
meristem growth	1.00E+00	1.00E+00	1.00E+00	1.00E+00	2.72E-02	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
embryonic meristem development	1.00E+00	1.00E+00	1.00E+00	2.07E-02	2.72E-02	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
polysaccharide biosynthetic process	1.00E+00	1.00E+00	1.00E+00	1.00E+00	3.05E-02	1.00E+00	2.10E-02	1.00E+00	1.00E+00	1.00E+00
secondary shoot formation	1.00E+00	1.00E+00	1.00E+00	1.00E+00	3.56E-02	1.00E+00	3.08E-02	1.00E+00	1.00E+00	1.00E+00
zinc II ion transport	1.00E+00	1.00E+00	1.00E+00	1.00E+00	3.87E-02	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
response to gibberellin	1.00E+00	1.00E+00	1.00E+00	1.00E+00	4.21E-02	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
potassium ion homeostasis	1.00E+00	1.00E+00	1.00E+00	1.00E+00	4.25E-02	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
threonine metabolic process	1.00E+00	1.00E+00	1.00E+00	1.00E+00	4.68E-02	1.00E+00	4.11E-02	1.00E+00	1.00E+00	1.00E+00
photosystem II repair	1.00E+00	1.00E+00	1.00E+00	1.00E+00	4.71E-02	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00

galactolipid metabolic process	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	2.23E-03	1.00E+00	1.00E+00
phosphatidylcholine metabolic process	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
glycolipid biosynthetic process	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	6.50E-03	1.00E+00	1.00E+00
response to red light	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
flower morphogenesis	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
regulation of ARF protein signal transduction	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
stomatal movement	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
cortical microtubule organization	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
lignin metabolic process	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
regulation of cell death	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
regulation of reactive oxygen species metabolic process	1.00E+00	1.00E+00	1.00E+00	2.06E-02	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
regulation of programmed cell death	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
zinc II ion transmembrane transport	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
cytoplasmic microtubule organization	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
isoprenoid metabolic process	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
cortical cytoskeleton organization	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
protein deubiquitination	1.00E+00	1.00E+00	2.54E-03	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
secretion by cell	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
meristem initiation	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
vesicle-mediated transport	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
peptidyl-tyrosine dephosphorylation	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
sulfate assimilation	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	8.41E-05	1.00E+00	5.94E-04	2.83E-07	1.00E+00	1.00E+00
endosperm development	1.00E+00	1.00E+00	1.00E+00	1.73E-02	1.00E+00	3.02E-04	1.00E+00	4.35E-04	5.24E-04	1.00E+00	1.00E+00
cotyledon development	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	6.03E-04	1.00E+00	1.63E-03	1.12E-03	1.62E-03	1.00E+00
embryo development ending in seed dormancy	1.00E+00	8.05E-05	6.45E-09	1.00E+00	1.00E+00	6.10E-04	1.00E+00	1.54E-10	3.45E-03	7.88E-07	1.00E+00
embryo development	1.00E+00	4.69E-05	7.56E-06	1.00E+00	1.00E+00	8.29E-04	1.00E+00	5.40E-12	2.75E-04	1.10E-06	1.00E+00
oxidoreduction coenzyme metabolic process	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.37E-03	1.19E-03	1.00E+00	1.00E+00	1.00E+00	1.00E+00
regulation of catalytic activity	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	3.88E-03	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
positive regulation of catalytic activity	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	4.10E-03	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
sucrose metabolic process	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	5.93E-03	1.00E+00	2.59E-04	2.75E-04	1.48E-04	1.00E+00

vacuolar transport	3.85E-02	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
coumarin biosynthetic process	3.87E-02	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
sulfur amino acid metabolic process	1.00E+00	2.51E-02	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
cellular amino acid biosynthetic process	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
sulfur compound biosynthetic process	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
regulation of gene silencing	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
serine family amino acid metabolic process	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
regulation of plant-type hypersensitive response	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
programmed cell death	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
protein targeting to membrane	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
salicylic acid mediated signaling pathway	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
cysteine biosynthetic process	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
cysteine metabolic process	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
cell death	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
systemic acquired resistance	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
sulfur compound metabolic process	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
cell redox homeostasis	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
RNA processing	1.00E+00	1.00E+00	1.00E+00	9.63E-04	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
specification of symmetry	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
regulation of immune response	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
protein autophosphorylation	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
determination of bilateral symmetry	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
host programmed cell death induced by symbiont	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
chromatin silencing by small RNA	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
regulation of innate immune response	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
brassinosteroid mediated signaling pathway	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	4.79E-02	1.00E+00	1.00E+00	1.00E+00
plant-type hypersensitive response	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
triterpenoid biosynthetic process	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
cellular response to brassinosteroid stimulus	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	4.31E-02	1.00E+00	1.00E+00	1.00E+00

cell wall macromolecule metabolic process	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
response to UV-B	1.00E+00	1.06E-09	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
flavonoid metabolic process	1.00E+00	8.11E-06	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
flavonoid biosynthetic process	1.00E+00	1.09E-05	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
pollen tube guidance	1.00E+00	2.31E-04	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
base-excision repair	1.00E+00	2.38E-04	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
response to sucrose	1.00E+00	2.45E-04	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
cytokinin biosynthetic process	1.00E+00	1.42E-03	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	4.53E-03
response to UV	1.00E+00	4.48E-03	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
maintenance of protein location	1.00E+00	6.96E-03	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
amino acid transmembrane transport	1.00E+00	2.04E-02	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
cytokinin metabolic process	1.00E+00	2.58E-02	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.48E-02
pollen sperm cell differentiation	1.00E+00	2.98E-02	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
proteasome-mediated ubiquitin-dependent protein catabolic process	1.00E+00	3.43E-02	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
chromatin assembly or disassembly	1.00E+00	4.63E-02	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
pentacyclic triterpenoid biosynthetic process	1.00E+00	4.68E-02	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
isocitrate metabolic process	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
regulation of transport	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
sister chromatid cohesion	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
regulation of proton transport	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
regulation of ion transport	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
nucleoside transport	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
nucleic acid phosphodiester bond hydrolysis	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
defense response signaling pathway, resistance gene-dependent	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
phosphate ion transport	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
chromosome segregation	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
meiotic chromosome segregation	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
DNA packaging	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	2.40E-03	1.00E+00	1.00E+00	1.00E+00	1.00E+00
nucleotide-sugar biosynthetic process	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	2.60E-03	1.00E+00	1.00E+00	1.00E+00	1.00E+00
nucleotide-sugar metabolic process	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	7.90E-03	1.00E+00	1.00E+00	1.00E+00	1.00E+00

										<u>Dominance deviation d</u>	
<u>DTE</u>	<u>LTE</u>	<u>Rosette</u> <u>dry mass</u>	<u>Rosette</u> <u>diameter</u>	<u>Area (day</u> <u>21)</u>	<u>Area (day</u> <u>29)</u>	<u>Perimeter</u> <u>(day 21)</u>	<u>Perimeter</u> <u>(day 29)</u>	<u>Area</u> <u>growth</u>	<u>Perimeter</u> <u>growth</u>		
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.08E-02	9.54E-04	7.08E-05	2.31E-05	2.32E-03		
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	9.12E-06	4.42E-02	1.84E-06	5.36E-06	1.46E-04		
1.00E+00	1.00E+00	1.00E+00	1.03E-04	3.68E-04	1.00E+00	8.78E-05	1.00E+00	1.00E+00	1.00E+00		
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.94E-02	1.00E+00	2.76E-02	1.62E-02	2.28E-02		
1.00E+00	1.00E+00	1.00E+00	1.00E+00	5.53E-05	7.67E-05	2.45E-05	5.71E-04	1.14E-04	2.55E-04		
1.00E+00	1.00E+00	1.00E+00	1.00E+00	6.67E-07	5.01E-03	9.64E-10	1.46E-04	2.75E-03	1.63E-04		
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	2.34E-02	1.00E+00	5.92E-03	3.05E-02	1.00E+00		
1.00E+00	1.00E+00	1.00E+00	4.56E-02	1.21E-06	7.34E-03	2.39E-09	2.48E-04	3.12E-03	1.99E-04		
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	9.49E-06	1.00E+00	3.77E-06	1.52E-07	4.79E-06		
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.74E-04	1.00E+00	4.12E-03	2.32E-04	2.72E-05		
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.40E-02	2.09E-03	4.81E-03		
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.32E-04	2.44E-04	1.08E-04	4.24E-04	1.70E-04	2.43E-04		
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	8.42E-04	1.00E+00	1.00E+00	1.00E+00		
1.00E+00	1.00E+00	1.00E+00	1.00E+00	4.04E-04	1.27E-03	3.79E-04	8.72E-04	1.16E-03	3.20E-04		
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.32E-02	1.00E+00	4.36E-04	8.14E-06	9.45E-05		
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	2.92E-03	1.00E+00	7.53E-03	1.50E-03	4.35E-03		
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	2.44E-04	1.00E+00	1.15E-03	2.60E-03	1.00E+00		
1.00E+00	1.00E+00	1.00E+00	6.13E-06	1.71E-04	1.00E+00	2.62E-03	1.00E+00	1.00E+00	1.00E+00		
1.00E+00	1.00E+00	1.00E+00	1.98E-05	7.30E-04	1.00E+00	7.56E-04	1.00E+00	1.00E+00	1.00E+00		
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	3.47E-02	1.00E+00	1.00E+00	1.00E+00		
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.32E-03	1.00E+00	9.41E-07	1.00E+00	1.00E+00	1.00E+00		
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	2.37E-02	1.00E+00	1.03E-02	4.34E-02	1.00E+00		
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.59E-03	1.56E-03	1.65E-03	1.34E-03	1.33E-03	1.17E-03		
1.00E+00	1.00E+00	1.00E+00	1.00E+00	2.14E-03	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00		
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	5.61E-05	1.00E+00	4.95E-04	4.31E-04	1.00E+00		
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.74E-04	1.00E+00	5.64E-04	1.38E-03	1.00E+00		

1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	6.85E-06	1.00E+00	2.27E-06	2.65E-06	7.97E-07
1.00E+00	1.00E+00	1.00E+00	1.00E+00	6.40E-09	3.58E-02	4.80E-07	1.00E+00	1.00E+00	1.00E+00
1.00E+00	1.00E+00	1.00E+00	1.00E+00	2.18E-07	1.00E+00	9.73E-06	1.00E+00	1.00E+00	1.00E+00
1.00E+00	1.00E+00	1.00E+00	1.00E+00	7.53E-07	1.00E+00	2.69E-05	1.00E+00	1.00E+00	1.00E+00
1.00E+00	1.00E+00	1.00E+00	8.55E-06	8.89E-06	1.00E+00	8.67E-06	1.00E+00	1.00E+00	1.00E+00
1.00E+00	1.00E+00	1.00E+00	1.00E+00	9.41E-05	1.00E+00	9.43E-03	1.00E+00	1.00E+00	1.00E+00
1.00E+00	1.00E+00	1.00E+00	1.00E+00	9.54E-05	1.00E+00	6.97E-06	1.39E-02	1.00E+00	1.00E+00
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.32E-04	1.00E+00	1.49E-05	1.78E-02	1.00E+00	1.00E+00
1.00E+00	1.00E+00	1.00E+00	1.00E+00	2.27E-04	1.00E+00	2.96E-02	1.00E+00	1.00E+00	1.00E+00
1.00E+00	1.00E+00	1.00E+00	1.00E+00	4.04E-04	1.00E+00	3.33E-03	8.27E-03	1.00E+00	1.00E+00
1.00E+00	1.00E+00	1.00E+00	1.00E+00	4.19E-04	1.00E+00	3.97E-04	1.00E+00	1.00E+00	1.00E+00
1.00E+00	1.00E+00	5.87E-03	1.00E+00	4.19E-04	1.00E+00	7.91E-03	5.80E-03	1.00E+00	1.00E+00
1.00E+00	1.00E+00	1.00E+00	1.00E+00	5.73E-04	5.89E-04	5.95E-04	5.07E-04	5.36E-04	1.00E+00
1.00E+00	1.00E+00	1.00E+00	1.00E+00	5.73E-04	6.03E-03	1.00E+00	1.00E+00	1.00E+00	1.00E+00
1.00E+00	1.00E+00	1.00E+00	1.00E+00	8.88E-04	1.00E+00	3.65E-02	1.00E+00	1.00E+00	1.00E+00
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.22E-03	1.00E+00	1.78E-05	2.50E-02	1.00E+00	1.00E+00
1.00E+00	1.00E+00	1.00E+00	1.95E-02	1.42E-03	1.00E+00	1.22E-02	1.00E+00	1.00E+00	1.00E+00
1.00E+00	1.00E+00	1.00E+00	1.00E+00	2.47E-03	2.68E-02	1.00E+00	1.00E+00	1.00E+00	1.00E+00
1.00E+00	1.00E+00	1.00E+00	1.00E+00	3.15E-03	1.00E+00	3.33E-03	1.00E+00	1.00E+00	1.00E+00
1.00E+00	1.00E+00	1.00E+00	1.00E+00	3.90E-03	1.00E+00	1.00E+00	2.09E-02	1.00E+00	1.00E+00
3.08E-04	1.00E+00	1.00E+00	1.00E+00	4.79E-03	1.00E+00	3.42E-02	1.00E+00	1.00E+00	1.00E+00
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.09E-02	1.00E+00	3.18E-02	1.00E+00	1.00E+00	1.00E+00
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.15E-02	1.00E+00	4.42E-02	1.00E+00	1.00E+00	1.00E+00
3.77E-04	1.00E+00	1.00E+00	1.00E+00	1.15E-02	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
1.00E+00	3.95E-02	1.00E+00	1.00E+00	1.27E-02	2.92E-03	1.26E-02	1.01E-02	3.12E-03	2.58E-03
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.39E-02	1.00E+00	3.18E-02	1.00E+00	1.00E+00	1.00E+00
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.42E-02	1.93E-04	1.50E-02	1.75E-04	1.79E-04	1.26E-04
1.00E+00	1.00E+00	8.85E-04	1.00E+00	1.43E-02	1.00E+00	1.00E+00	1.00E+00	1.00E+00	3.77E-02
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.59E-02	1.00E+00	1.00E+00	1.06E-03	1.00E+00	1.00E+00
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.60E-02	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00

1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.61E-02	1.00E+00	1.00E+00	7.72E-04	1.00E+00	1.00E+00
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.67E-02	1.00E+00	1.85E-02	1.00E+00	1.00E+00	1.00E+00
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.69E-02	1.00E+00	1.00E+00	2.57E-03	1.00E+00	1.00E+00
1.00E+00	1.00E+00	1.00E+00	1.00E+00	2.07E-02	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
4.30E-02	1.00E+00	1.00E+00	1.00E+00	2.20E-02	2.75E-02	1.00E+00	2.54E-02	1.00E+00	1.00E+00
1.00E+00	1.00E+00	1.00E+00	1.00E+00	2.29E-02	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
1.00E+00	1.00E+00	1.00E+00	1.00E+00	2.61E-02	1.00E+00	1.00E+00	2.71E-04	1.00E+00	1.00E+00
1.00E+00	1.00E+00	1.00E+00	1.70E-03	2.68E-02	1.00E+00	1.00E+00	9.45E-12	1.00E-06	7.13E-08
1.00E+00	1.00E+00	1.00E+00	1.00E+00	2.96E-02	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
5.03E-04	1.00E+00	1.00E+00	1.00E+00	3.23E-02	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
1.00E+00	1.00E+00	1.00E+00	2.07E-03	3.25E-02	1.05E-04	1.00E+00	3.73E-04	5.85E-05	2.86E-03
3.94E-04	1.00E+00	1.00E+00	1.00E+00	3.50E-02	4.87E-02	1.00E+00	1.00E+00	1.00E+00	1.00E+00
1.00E+00	1.00E+00	1.00E+00	1.00E+00	3.59E-02	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
1.00E+00	1.00E+00	1.00E+00	2.30E-03	3.79E-02	1.00E+00	1.00E+00	2.88E-11	1.82E-06	1.69E-07
1.00E+00	1.00E+00	1.00E+00	1.00E+00	4.00E-02	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
1.00E+00	1.00E+00	1.00E+00	3.20E-03	4.41E-02	1.00E+00	1.00E+00	6.70E-13	2.76E-06	3.18E-07
1.00E+00	1.00E+00	1.00E+00	1.00E+00	4.65E-02	1.00E+00	1.69E-05	1.00E+00	1.00E+00	1.00E+00
1.00E+00	1.00E+00	1.00E+00	1.00E+00	4.67E-02	9.35E-04	4.47E-02	1.01E-03	9.54E-04	7.47E-04
1.00E+00	1.00E+00	1.00E+00	1.00E+00	4.72E-02	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
1.00E+00	1.00E+00	1.00E+00	1.00E+00	4.72E-02	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
1.00E+00	1.00E+00	1.00E+00	1.00E+00	4.85E-02	1.00E+00	1.00E+00	1.00E+00	1.00E+00	2.83E-03
1.41E-04	1.00E+00	1.00E+00	1.00E+00	1.00E+00	9.76E-03	1.00E+00	1.00E+00	2.88E-07	1.00E+00
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.69E-03	1.00E+00	7.08E-05	3.71E-04	7.84E-05
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.50E-02	1.00E+00	1.12E-02	7.17E-03	2.93E-03
1.00E+00	1.00E+00	8.85E-04	1.00E+00	1.00E+00	1.00E+00	1.00E+00	5.02E-05	6.77E-03	6.40E-08
1.00E+00	1.00E+00	1.20E-02	1.00E+00	1.00E+00	1.00E+00	1.00E+00	6.42E-05	1.18E-02	2.50E-06
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	8.95E-04	1.00E+00	5.80E-03	5.04E-03	1.00E+00
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	3.19E-04	1.00E+00	1.00E+00
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	4.42E-03	1.00E+00	1.37E-02	1.00E+00	1.00E+00
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.69E-02	1.00E+00	1.78E-02	2.17E-05	1.05E-05

1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	3.69E-04	1.00E+00	1.00E+00	6.05E-03	1.00E+00
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	3.54E-04	1.00E+00	1.00E+00	6.98E-03	1.00E+00
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	5.63E-03	1.44E-02	1.36E-02
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	5.71E-04	1.00E+00	9.29E-04	2.92E-04	2.22E-04
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.86E-03	9.08E-03	2.85E-03
1.00E+00	1.00E+00	4.86E-02	1.00E+00	1.00E+00	1.38E-02	1.00E+00	3.81E-05	2.27E-02	3.37E-07
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	3.09E-02	1.00E+00	2.96E-02	2.73E-02	3.47E-03
1.00E+00	1.00E+00	1.28E-03	1.00E+00	1.00E+00	1.32E-02	1.00E+00	5.50E-05	2.36E-02	3.18E-07
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.83E-07	1.00E+00	2.65E-09	4.89E-08	6.26E-07
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	3.47E-06	1.00E+00	3.03E-03	1.18E-06	1.36E-02
1.00E+00	1.00E+00	1.00E+00	3.00E-02	1.00E+00	4.49E-06	1.00E+00	1.46E-07	1.32E-07	1.05E-05
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	5.42E-06	1.00E+00	1.59E-03	2.76E-06	1.00E+00
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	2.16E-05	1.00E+00	9.54E-05	1.38E-05	7.20E-04
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	4.80E-04	1.00E+00	4.93E-04	7.40E-05	1.00E+00
1.00E+00	1.00E+00	3.86E-02	1.00E+00	1.00E+00	1.69E-03	1.00E+00	1.00E+00	8.69E-04	1.00E+00
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	2.41E-03	1.00E+00	6.29E-03	7.72E-04	3.58E-04
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	2.92E-03	1.00E+00	8.90E-05	2.24E-03	1.34E-02
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	3.21E-03	1.59E-02	4.88E-03	5.54E-04	2.64E-02
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	3.55E-03	1.00E+00	6.95E-05	2.17E-08	3.48E-07
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	4.05E-03	1.57E-03	2.90E-02	4.32E-03	1.00E+00
8.57E-03	1.00E+00	1.00E+00	1.00E+00	1.00E+00	5.17E-03	1.00E+00	3.04E-02	2.50E-03	1.00E+00
6.97E-04	1.00E+00	1.00E+00	1.00E+00	1.00E+00	5.67E-03	1.00E+00	2.43E-02	2.70E-02	1.00E+00
4.06E-02	1.00E+00	1.00E+00	1.00E+00	1.00E+00	7.86E-03	1.00E+00	8.02E-03	5.47E-03	3.47E-02

2.18E-03	2.22E-03	1.00E+00	1.00E+00	1.00E+00	7.86E-03	1.00E+00	4.35E-02	4.13E-02	1.00E+00
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.26E-02	1.00E+00	3.42E-04	9.63E-05	4.50E-04
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.28E-02	1.00E+00	2.21E-08	1.11E-03	7.13E-08
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.34E-02	1.00E+00	8.20E-04	1.36E-02	6.44E-04
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.40E-02	1.00E+00	4.57E-02	1.00E+00	1.00E+00
1.00E+00	2.79E-02	1.00E+00	1.00E+00	1.00E+00	1.69E-02	1.00E+00	1.00E+00	1.90E-02	1.79E-02
1.00E+00	1.00E+00	1.00E+00	1.02E-06	1.00E+00	1.69E-02	1.00E+00	1.00E+00	1.61E-02	1.00E+00
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	2.75E-02	1.00E+00	2.57E-03	2.62E-02	1.00E+00
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	3.52E-02	1.00E+00	2.63E-07	3.67E-03	3.18E-07
1.00E+00	1.00E+00	1.61E-03	1.00E+00	1.00E+00	3.59E-02	1.00E+00	3.98E-05	2.26E-03	8.97E-03
1.58E-02	1.00E+00	1.00E+00	1.00E+00	1.00E+00	4.17E-02	1.00E+00	1.00E+00	1.00E+00	1.00E+00
1.00E+00	1.00E+00	4.88E-03	3.59E-02	1.00E+00	4.44E-02	1.00E+00	4.17E-02	1.00E+00	1.00E+00
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	4.56E-02	1.00E+00	8.68E-05	1.23E-04	9.16E-05
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	4.91E-02	1.00E+00	1.00E+00	4.65E-02	1.00E+00
1.00E+00	1.00E+00	1.00E+00	1.67E-03	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.10E-04	5.97E-05
1.00E+00	1.00E+00	1.00E+00	4.18E-03	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.29E-04	9.78E-05
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
1.00E+00	1.00E+00	1.28E-03	3.13E-02	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.15E-02	1.00E+00
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
3.92E-02	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	6.10E-04	2.26E-07	7.72E-10
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	3.18E-02	2.74E-05	2.74E-03
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	3.43E-02	3.45E-05	3.38E-03
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	3.70E-03	3.81E-05	3.36E-03
1.00E+00	2.73E-02	1.22E-05	1.00E+00	1.00E+00	1.00E+00	1.00E+00	5.33E-04	3.82E-05	3.42E-04
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	3.99E-05	1.00E+00
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	2.17E-04	4.33E-05	8.38E-08
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	3.03E-04	2.26E-04
1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	5.15E-03	7.36E-04	2.24E-02

Table S9. Genomic control (GC) per GWA experiment.

	<u>Additive encoding</u>	<u>Overdominant encoding</u>	
	<u>Predicted mean phenotype</u>		<u>Dominance deviation <i>d</i></u>
DTF	1.00	0.97	0.95
LTF	1.00	0.97	0.95
Dry mass	1.03	0.97	0.97
Rosette diameter	1.11	1.01	0.97
Area (day 21)	1.08	1.01	1.03
Perimeter (day 21)	1.09	0.99	1.03
Area (day 29)	1.13	0.99	1.01
Perimeter (day 29)	1.12	1.02	1.01
Area growth	1.14	0.98	1.00
Perimeter growth	0.95	0.98	0.99

```

/*Import data*/
FILENAME hybinput '/folders/myfolders/
2015-04-06_sas_hybrids_transformed_all_continuous_pheno.txt';
DATA hybdata;
    INFILE hybinput DELIMITER=', ' firstobs=2;
    INPUT female male genotype ltf dtf rosdiam drymass area21
perim21 area29 perim29 areagrowth perimgrowth;
RUN;

/*Create Z matrix for random effects in model*/

/*Macro variable "dlset"*/
%LET dlset=hybdata;
%LET tempvar=perimgrowth; *change here to analyze a different
phenotype;

/*sort data and create list of parents "plist"*/
PROC SORT DATA=&dlset;
    BY female male;

/*Create a list of males and females*/
PROC SUMMARY DATA=&dlset NOPRINT;
    CLASS female male;
    OUTPUT OUT=plist(where=( _type_=3));

TITLE 'List of females, males, and number of samples per cross';
PROC PRINT DATA=plist noobs;
    VAR female male _FREQ_;
RUN;

/*Combine parents into single variable*/
DATA parent;
    SET plist(rename=(female=parent))
        plist(rename=(male=parent));

PROC SUMMARY DATA=parent(keep=parent);
    CLASS parent;
    OUTPUT OUT=parent(where=( _type_=1));

DATA parent(drop=_type_ _freq_ pn);
    SET parent;
    pn+1;
    CALL SYMPUT('pn',compress(pn));

TITLE 'List of parents';
PROC PRINT DATA=parent;
RUN;

/*construct dummy variables p1-p31d*/
PROC IML;

```

```

USE &dlset;
READ ALL VAR {female male} INTO d;
CLOSE &dlset;
n=NROW(d);

*create matrix (pn x4) with parent, parent code (1-pn);
USE parent;
READ ALL VAR {parent} into p;
CLOSE parent;
p=CHAR(p); *I had to add this line - without it I got an error
at line 61;
pcode=CHAR(1:NROW(p),5,0)`;

*create pcode corresponding to parent code in dummy;
p=p||pcode;
PRINT n p; *Checks number of observations and number of
parents;

CREATE pcode FROM pcode [COLNAME={'p'}];
APPEND FROM pcode;
CLOSE pcode;

*create dummy variables;
a=SHAPE(0,n,NROW(p));
DO i=1 to n;
    DO k=1 to nrow(p);
        IF CHAR(d[i,1])=p[k,1] | CHAR(d[i,2])=p[k,1]
then a[i,k]=1;
    END;
END;
CREATE dummy from a;
APPEND FROM a;
CLOSE dummy;
QUIT;

/*Merge dummy variables with original data*/
DATA &dlset;
    MERGE &dlset dummy;
PROC SORT DATA=&dlset;
    BY genotype;
RUN;

TITLE 'Data with dummy variables;';
PROC PRINT DATA=&dlset (OBS=10) NOOBS;
VAR female male genotype &tempvar col1-col30;
RUN;

/*Run PROC MIXED on variable of choice*/
PROC MIXED DATA=&dlset COVTEST ASYCOV UPDATE;
CLASS genotype;

```

```

        MODEL &tempvar= / solution;
        RANDOM col1-col&n/TYPE=TOEP(1) solution; *GCA effects -
solution keeps track of BLUP;
        RANDOM genotype; *SCA effects;
ODS OUTPUT COVPARMS=_varcomp ASYCOV=_cov solutionF=_intercept
solutionR=_BLUPvar;
RUN;

```

```

PROC EXPORT DATA=_intercept OUTFILE="/folders/myfolders/
&tempvar._intercept_SAS.txt"
DBMS=tab replace;
RUN;

```

```

PROC EXPORT DATA=_BLUPvar (obs=30) OUTFILE="/folders/myfolders/
&tempvar._BLUP_SAS.txt"
DBMS=tab replace;
RUN;

```

```

/*Estimate additive genetic variance by extracting varcomp from the
initial run*/

```

```

PROC IML;
/*Create column vector of variance components*/
USE _varcomp;
READ all var {Estimate} into VC;
CLOSE _varcomp;

```

```

/*Create matrix of covariances of variance components*/
USE _cov;
READ all var {CovP1 CovP2 CovP3} into COV;
CLOSE _cov;

```

```

/*Create vector of coefficients for the numerator of narrow sense h2*/
AU=SHAPE(0,nrow(VC),1);
AU[1,1]=1*2;

```

```

/*Create vector of coefficients for the numerator of broad sense h2*/
AD=SHAPE(0,nrow(VC),1);
AD[1,1]=1*2;
AD[2,1]=1*1;

```

```

/*Create vector of coefficients for the denominator of h2*/
AV=SHAPE(1,nrow(VC),1);
AV[1,1]=2;

```

```

/*Calculate phenotypic variance*/
Total=VC[+,1]; *Sum of VC column - total observed variance;
Pheno=AV`*VC; *Phenotypic variance (2*GCA + SCA + resid);

```

```

/*Calculate broad and narrow sense h2*/
h2n=AU`*VC/Pheno; *Additive/Phenotypic;
h2b=AD`*VC/Pheno; *Additive+dominance/Phenotypic;

/*Calculate the percent of variances by each term*/
VC_pct=VC/Total*100; *Percentage of variance by term;
Var_VC=VECDIAG(COV); *Variance of variances;
SE_VC=sqrt(Var_VC); *Standard errors of variances;

/*Use the delta method to calculate the standard error of each h2*/
var_U=AU`*Cov*AU; *variance of additive numerator;
var_D=AD`*Cov*AD; *variance of additive+dominance numerator;
var_V=AV`*Cov*AV; *variance of denominator;

cov_UV=AU`*Cov*AV; *covariance between variances;
cov_DV=AD`*Cov*AV; *covariance between variances;

SEh2n=sqrt((h2n*h2n)*((var_U/(AU`*VC)**2)+(var_V/(AV`*VC)**2)-
(2*cov_UV/(AU`*VC)/(AV`*VC))));
SEh2b=sqrt((h2b*h2b)*((var_D/(AD`*VC)**2)+(var_V/(AV`*VC)**2)-
(2*cov_DV/(AD`*VC)/(AV`*VC))));

/*Create empty table to store results and print out results*/
tempcol = SHAPE(0,7,1);
h2stats = tempcol||tempcol||tempcol;
names = {VC VCper PhenoVar h2n h2b seh2n seh2b};
h2stats[1,1:3]=VC`;
h2stats[2,1:3]=VC_pct`;
h2stats[3,1]=Pheno;
h2stats[4,1]=h2n;
h2stats[5,1]=h2b;
h2stats[6,1]=seh2n;
h2stats[7,1]=seh2b;

CREATE h2out FROM h2stats[ROWNAME=names];
APPEND FROM h2stats[ROWNAME=names];

RUN;

PROC EXPORT DATA=h2out OUTFILE="/folders/myfolders/
&tempvar._h2stats_SAS.txt"
DBMS=tab replace;
PUTNAMES=yes;
RUN;

QUIT;

```

6. “Recurrent segregation distortion is uncovered in a species-wide screen for biased genetic transmission”

Seymour DK, Chae E, Ariöz B, Koenig D, Weigel D. (In preparation).

Abstract

The equal probability of inheritance of alleles during sexual reproduction is a central tenet of genetics and evolutionary biology. Yet, there are many cases where this rule is violated. Such violations limit intraspecific gene flow and can facilitate the formation of genetic barriers, a first step in speciation. Biased transmission of alleles, or segregation distortion, can result from a number of biological processes including epistatic interactions between incompatible loci, gametic selection, and meiotic drive. Examples of these phenomena have been identified in many species implying that they are universal, but comprehensive species-wide studies of segregation distortion have not yet been undertaken. We have performed a species-wide screen for distorted allele frequencies in over five hundred F_2 populations using reduced-representation high-throughput sequencing. We have found that the biased transmission of alleles is prevalent, occurring in 12-24% of surveyed populations. Additionally, we find that most populations exhibit distortion at only one genomic region and that some regions are repeatedly distorted in multiple populations. This data set elucidates the species-level distribution of biased transmission of genetic material in *A. thaliana*, and serves as a springboard for studies of the genetic basis of intraspecific genetic barriers.

Contributions

Conceived and designed the experiments: DKS DK EC DW. Performed the experiments: DKS BA. Analyzed the data: DKS. Contributed to the writing of the manuscript: DKS DW.

1 **Title**

2 **Recurrent segregation distortion is uncovered in a species-wide screen for biased**
3 **genetic transmission**

4

5 **Authors**

6 Danelle K. Seymour^{1,2}, Eunyoung Chae¹, Burak Ariöz¹, Daniel Koenig^{1,3}, Detlef Weigel^{1*}

7

8 **Affiliations**

9 ¹Department of Molecular Biology, Max Planck Institute for Developmental Biology, 72076
10 Tübingen, Germany

11 ²Current address: Department of Ecology and Evolutionary Biology, University of
12 California, Irvine, CA, USA

13 ³Current address: Department of Botany and Plant Sciences, University of California,
14 Riverside, CA, USA

15 *Corresponding author, email: weigel@weigelworld.org

16 **Abstract**

17 The equal probability of inheritance of alleles during sexual reproduction is a central tenet
18 of genetics and evolutionary biology. Yet, there are many cases where this rule is violated.
19 Such violations limit intraspecific gene flow and can facilitate the formation of genetic
20 barriers, a first step in speciation. Biased transmission of alleles, or segregation distortion,
21 can result from a number of biological processes including epistatic interactions between
22 incompatible loci, gametic selection, and meiotic drive. Examples of these phenomena
23 have been identified in many species implying that they are universal, but comprehensive
24 species-wide studies of segregation distortion are lacking. We have performed a species-
25 wide screen for distorted allele frequencies in over five hundred segregating populations
26 using reduced-representation high-throughput sequencing. We have found that the
27 biased transmission of alleles is evident in 12-24% of surveyed populations. Additionally,
28 we find that most populations exhibit detectable distortion at only one genomic region and
29 that some regions are repeatedly distorted in multiple populations. This data set
30 elucidates the species-level distribution of biased transmission of genetic material in *A.*
31 *thaliana*, and serves as a springboard for future studies of the genetic basis of intraspecific
32 genetic barriers.

33

34 **Introduction**

35 At the genetic level, evolution is the change in the frequency of allelic variants over
36 time. While in many cases, the strength of selection is too low for these changes to be
37 detected within a few generations, a unique opportunity to directly study such changes is
38 offered in cases where selection coefficients are high. In such a situation, competition
39 between alleles can be seen already in the progeny of heterozygous (*a/b*) plants or
40 animals. It is manifested as a deviation from the 1:2:1 Mendelian ratio of diploid genotypes
41 (*a/a*, *a/b*, *b/b*), termed allelic or segregation distortion. Deviation from this ratio has
42 important implications for population dynamics. Favoring inheritance of an allele from one
43 grandparent over that from the other grandparent implies that certain genotypic
44 combinations may be unfit [1-3]. Depending on the underlying mechanism, this
45 obstruction to the free flow of genetic information is an irreversible step on the path
46 towards speciation (reviewed in [4]).

47 Segregation distortion, which is quite commonly observed in nature, can be the
48 result of deleterious epistatic interactions between incompatible loci, gametic selection,
49 or meiotic drive (reviewed in [5]). Perhaps epistatic interactions of the Bateson-
50 Dobzhansky-Muller type are the most well studied examples [6]. Alone, incompatible
51 mutations are innocuous in their native genetic environment. But, when combined,
52 reduced fitness or lethality removes incompatible genotypic combinations from the
53 population. Examples of two-locus incompatibilities have been identified in and between
54 several eukaryotic species (reviewed in [7, 8]) and causal loci are frequently associated
55 with fast molecular evolution (reviewed in [8]). Likely, epistatic incompatibilities are a
56 common topic in the literature not only due to their role in speciation, but also because
57 they are relatively easy to detect.

58 Far fewer examples of meiotic drive and gametic selection have been
59 characterized. Meiotic drive refers to the preferential inheritance of one chromosome
60 during meiosis and is most easily discovered during female gametogenesis [9], as only
61 one of the four meiotic products will become the egg nucleus. This creates the opportunity
62 for “selfish” loci to position their nucleus favorably so that their transmission is
63 favored in the next generation. Some known examples of female drive involve changes
64 in either centromeric or other heterochromatic regions [1, 10], possibly favoring
65 transmission of the drive chromosome by increasing its affinity for the meiotic machinery
66 [1, 9, 11-16]. Many of the drive loci are located on sex chromosomes (esp. in *Drosophila*)
67 and are associated with inversions or other cytological changes [1, 11, 16]. Drive loci on
68 sex chromosomes are more readily identified because they alter the sex ratio, which is
69 easily detected without molecular genotyping.

70 Transmission biases arising after formation of the haploid gametes are classified
71 as instances of gametic selection. Due to the differences of male and female
72 gametogenesis, gametic selection can be more easily detected in males. Sperm is
73 produced from all four meiotic products, and each of these haploid sperm cells can
74 compete for the ability to fertilize the ovule. A classic example of gametic selection
75 involves growth of the pollen tubes that deliver the male gametes of plants [17].
76 Differential pollen tube growth can improve the reproductive success of the genotype that
77 elongates more quickly. Competition among haploid sperm results in biased transmission

78 of the more fit allele in the next generation.

79 Though a few instances of segregation distortion are relatively well studied,
80 comprehensive species-wide studies of biased transmission are limited. Despite the
81 apparent ubiquity of segregation distortion, it is unclear how often epistatic
82 incompatibilities, gametic selection, or meiotic drive are the cause. In *A. thaliana*,
83 segregation distortion has been observed repeatedly in many experimental population
84 designs [18-20]. To date, the largest study in *A. thaliana* examined segregation distortion
85 in 17 F₂ populations, over half of which exhibited evidence of distortion [20]. Although *A.*
86 *thaliana* is typically a self-fertilizing species, outcrossing in nature can be quite common,
87 implying that opportunities for unequal transmission exist [21]. Preference for inbreeding
88 creates a system sensitized for detection of intraspecific distortion. Cross-fertilization of
89 accessions removes an allele from its native, homozygous context, thus creating the
90 opportunity for biased transmission, making *A. thaliana* an ideal system for the
91 identification of preferentially inherited loci.

92 We surveyed over 500 segregating F₂ populations for segregation distortion in
93 order to characterize the contribution of biased transmission to the generation of
94 intraspecific genetic barriers. Segregating F₂ populations were derived from intercrossing
95 80 distinct *A. thaliana* accessions representing the known Eurasian genetic diversity of
96 the species [22]. For this large survey, populations were genotyped in pools using
97 reduced-representation high-throughout sequencing to estimate allelic ratios. In addition
98 to characterizing the prevalence of segregation distortion in *A. thaliana*, we also sought
99 to dissect the genetic architecture underlying biased loci. Detailed molecular
100 characterization is will help to determine the relative contribution of deleterious epistatic
101 interactions, male gametic selection, or female drive meiotic to biased inheritance.

102

103 **Results**

104 *Segregation distortion arises frequently in intraspecific A. thaliana F₂ populations*

105 The incidence of segregation distortion, a molecular signature of genetic conflict, was
106 surveyed in 583 F₂ populations generated from germplasm that represents the known
107 genetic diversity in *A. thaliana* [22]. The studied F₂ populations were derived from crosses
108 between 67 female and 80 male grandparental accessions with sequenced genomes [22].

109 The number of crosses performed per accession ranged from 3 to 34, with a median of
110 14 F₂ populations generated from each grandparent.

111 Restriction enzyme-mediated reduced-representation sequencing (RAD-Seq)
112 facilitated the genotyping of a large number of populations. Based on previous reduced-
113 representation approaches [23, 24], a custom protocol was developed to tackle the
114 specifics of this work. Populations were genotyped in bulks consisting of at least 300 F₂
115 individuals and linkage information was derived from grandparental whole-genome data.
116 Accurate allele frequency estimate in bulks requires high coverage at each segregating
117 site. The selected restriction enzyme, KpnI, cuts infrequently in the *A. thaliana* genome
118 allowing high coverage to be achieved for a reduced portion of the genome (1%). We
119 attained an average coverage of 78X per F₂ population (Fig 1A) and an average of 2,509
120 sites were segregating in any given population (Fig 1B) close to what had been predicted
121 from the whole-genome resequencing data.

122 Regions displaying significant segregation distortion were identified by modeling
123 the allele frequency in 5 Mb sliding windows (0.5 Mb steps). Using the beta-binomial
124 model estimates of predicted allele frequency together with the confidence intervals of
125 that estimate, a non-parametric statistical test was performed in each window. In total, 62
126 populations exhibited regions of significant segregation distortion after false discovery
127 rate (FDR) correction for the number of tested genomic windows ($n = 240$, $p < 0.05$).
128 When considering only populations passing quality control measures ($n = 492$), 12.6% of
129 surveyed populations are significantly distorted (Fig 2). This is a rather conservative
130 estimate of the frequency of segregation distortion in our data set because the ability to
131 detect significant distortion is highly dependent on the size of the confidence interval
132 estimates (i.e. the coverage of each population). To generate a less conservative
133 estimate of the number of distorted regions we also used a Z-score outlier approach. Any
134 region with an allele frequency greater than 2.5 standard deviations from the combined
135 population mean was considered to be distorted. This less conservative approach
136 identified 122 populations with at least a single distorted region or 24.8% of the
137 populations surveyed (Fig 3). All regions identified via the FDR method were also
138 detected using the Z-score outlier approach. An example of a distorted chromosome
139 identified using both methods is shown in Figure 4. Although we did not screen the

140 complete diallel of possible F_2 combinations, we did survey populations across that
141 genetic space (Fig 2, Fig 3). That segregation distortion is evident in 12 to 24% of
142 surveyed F_2 populations suggests that intraspecific genetic barriers are much more
143 common than previously anticipated.

144

145 *The dynamics of segregation distortion in A. thaliana*

146 The underlying genetic basis of segregation distortion is dictated by the biological process
147 driving the observed non-Mendelian inheritance. To understand the relative contribution
148 of these processes, i.e. genetic incompatibility, meiotic drive, and gametic selection, we
149 further characterized the dynamics of segregation distortion in our data set. Regardless
150 of identification method (FDR or outlier), the majority of populations exhibit distortion at
151 only a single locus (Fig 5A). If classical Bateson-Dobzhansky-Muller genetic
152 incompatibilities were driving segregation distortion in our populations, we would expect
153 two distorted regions per population, unless such loci are typically linked. We also found
154 that distortion occurs on all five chromosomes, although distorted regions are most
155 frequently located on chromosome 1 (Fig 5B).

156 The alleles favored in distorted regions are derived from many grandparental
157 accessions. Of the 80 accessions, over 50 give rise to F_2 populations exhibiting significant
158 segregation distortion. Some grandparents are especially notable, Star-8 for example.
159 Star-8 derived alleles are distorted in 60% of F_2 populations from this accession (40% for
160 the FDR threshold) (Fig 6A,B).

161 If genetic barriers arise due to genetic drift, as individuals diverge from a common
162 ancestor we would expect more distantly related accessions to give rise to distortion more
163 frequently. As the grandparental accessions are representative of Eurasian genetic
164 diversity, we were able to test if increased genetic diversity between the two F_2
165 grandparents spawned segregation distortion. We found that there was no significant
166 difference between the grandparental genetic distance of distorted populations versus
167 that of non-distorted, or normal, populations at a 1% significance threshold ($p=0.03$
168 (outlier distortion), $p = 0.11$ (FDR distortion), Wilcoxon rank-sum test) (Fig 7A,B). That
169 genetic diversity is not predictive of segregation distortion suggests that non-stochastic
170 processes give rise to intraspecific genetic barriers in *A. thaliana*.

171

172 *Refining candidate intervals surrounding distorted loci*

173 Finally, we sought to characterize the genetic basis underlying distorted regions.
174 Genotyping populations bulks enabled screening of a large number of genetically diverse
175 segregating populations, but without genotype information from individual segregants to
176 estimate recombination breakpoints, candidate regions span almost entire chromosomes
177 arms. Since genotyping a large number of individuals from multiple distorted populations
178 was still cost prohibitive, we designed two strategies to narrow the candidate regions to
179 facilitate subsequent fine-mapping.

180 First, we generated whole-genome resequencing data for six populations
181 displaying severe segregation distortion. For each population, we sequenced pools of
182 1,500 F₂ individuals to approximately 50X coverage in order to further narrow the intervals
183 surrounding the candidate loci. Although the pool of recombination events increased, the
184 lower sequencing coverage was accompanied by increased stochasticity in allele
185 frequency estimates. We took advantage of local linkage disequilibrium to diminish that
186 noise. Short stretches of unique 21 nucleotide (nt) sequences (known as k-mers or 21-
187 mers) were identified in the raw sequencing reads of each pool. Any 21-mer sequence
188 shared between grandparents should occur at the average genome-wide coverage.
189 Sequences present in only one of the two parents should have approximately half as
190 much coverage. Peaks of 21-mer coverage at ~25X and ~50X are found in all six
191 populations (Fig 8). To narrow candidate intervals, we used a sliding window approach (1
192 Mb window, 50 kb step) to calculate the average coverage of 21-mers that occur in only
193 one of the grandparents. Regions of the genome that are distorted should display a
194 decrease in coverage near the causal locus. Using this strategy, we were able to narrow
195 the intervals surrounding four of the six candidate loci from chromosome arms to less
196 than 5 Mb, and in one cases the region was as small as 1.5 Mb (Table 2, Fig 9).

197 Instead of increasing the number of recombination events, the second approach
198 to refine candidate regions relied on obtaining a precise estimate of allele frequency by
199 massively increasing the sequencing coverage. As mentioned earlier, some
200 grandparental accessions contributed alleles that were favored in multiple F₂ populations.
201 Three accessions in particular, Star-8, ICE63, and ICE49, contributed alleles that were

202 favored in at least 40% of crosses derived from them (based on the outlier method) and,
203 in each case, the same regions were favored in all distorted populations sharing a
204 particular grandparent. Using a bulked segregant analysis (BSA) approach [25], we
205 generated two super pools of reads for each grandparent. One super pool comprised the
206 sequencing reads from distorted populations and the other contained the combined
207 sequencing reads from the normal populations. The allele frequency of SNPs was
208 calculated for sites segregating between the focal grandparent and all other accessions
209 in either the distorted bulk or the normal bulk. A median coverage of at least 806X was
210 achieved at each segregating site, vastly improving the accuracy of our estimation. For
211 one grandparent, Star-8, we were able to narrow the interval to 2.04 Mb (Table 2, Fig
212 10A). This strategy was less successful for the other two grandparents, likely because of
213 the decreased strength of distortion in these regions as well as the location of these loci
214 on the chromosome. The Star-8 locus resides in the middle of the top arm of chromosome
215 1 (Fig 10A), allowing recombination to occur on either side of it. The other loci were
216 located where recombination is often reduced, either near a centromere (Fig 10B) or on
217 a distal chromosome arm.

218

219 **Discussion**

220 Despite the ubiquity of non-Mendelian segregation of alleles in natural populations, the
221 genetic and molecular characterization of the responsible loci has been lagging (reviewed
222 in [1, 2, 5, 16, 26-28]). Such systems are most easily studied, when distortion is severe
223 and differences in phenotypically distinct progeny classes are obvious (reviewed in [16]).
224 Because sexual dimorphism is common, many of the earliest known cases were
225 discovered because sex-ratio deviated greatly from 1:1 (reviewed in [16]). The effects of
226 an allele that is preferentially inherited can be neutralized in a population by fixation of the
227 allele or by the evolution of secondary modifiers. Many cases of segregation distortion
228 were discovered in interspecific crosses [1, 29-33], not because the phenomenon is more
229 common in interspecific hybrids, but because the severity of distortion is extreme in the
230 absence of species-specific modifiers, sometimes reaching fixation in only a generation
231 or two [1]. The same loci responsible for segregation distortion in interspecific crosses
232 may also underlie unexpected intraspecific segregation patterns. However, in

233 intraspecific crosses, allele frequencies are often only perturbed by a few percent [1, 5],
234 and without molecular genotyping techniques, such subtle allelic distortion will go mostly
235 undetected.

236 Exploiting advances in sequencing and genotyping technology, we have been able
237 to characterize segregation distortion in hundreds of intraspecific crosses. The
238 identification of distorted regions greatly depends on sequencing coverage; in our system,
239 a 10% deviation in absolute allele frequency becomes significant with approximately 100x
240 sequence coverage, and more subtly distorted regions could be detected with even higher
241 coverage. Similar pooled genotyping approaches have been used to identify distorted loci
242 in other systems [34-37], illustrating the general power of this approach.

243 Although *A. thaliana* is self-compatible, outcrossing is reasonably common, and
244 descendants of recent outcrossing events are easily found in wild stands of this species
245 [21]. By surveying a broad collection of germplasm for non-Mendelian inheritance, we
246 could confirm that allelic distortion is a common feature of F₂ populations, implying that
247 allelic distortion has a major impact on shaping local genetic diversity. Not only do
248 distorted loci segregate in up to a quarter of all F₂ populations, but multiple genomic
249 regions contribute to this phenomenon, with the degree of distortion varying both by
250 population and by locus. Intraspecific distortion loci that have been identified in other
251 systems typically occur at low population frequencies [38-44], although there are
252 exceptions, such as the tightly linked *zeel-1* and *peel-1* genes in *C. elegans* [45, 46]. The
253 low frequency of the causal alleles has been hypothesized to result from antagonistic
254 modifier loci having evolved in response to the fitness costs that are often linked to
255 distortion loci (reviewed in [5, 16]). In an interspecific *Drosophila* cross, the causal locus
256 itself is responsible for both the distortion phenotype and for reduced gamete success [2].
257 We have found multiple cases of genomic regions that are distorted in one or very few
258 population(s), suggesting that frequency of distortion alleles is often low in *A. thaliana* as
259 well. This could be because these alleles are older, giving sufficient time for modifiers to
260 evolve and rise to high frequency. If these are linked, we would not have detected them
261 as separate genomic loci, as our mapping resolution was mostly chromosome arm scale.

262 Of particular interest are regions that are repeatedly distorted across many
263 populations at extreme frequencies. For example, the Star-8 region on chromosome 1 is

264 significantly favored in ~50% of crosses, with this region being inherited by up to 70 or
265 even 80% of the progeny. This could be an example of a young allele for which
266 suppressors have not yet evolved, or it could be that the balance between fitness costs
267 (if any) and the degree of distortion is stable at this frequency. The *D* locus in *Mimulus*
268 *guttatus* is perhaps the best example of a stable distortion polymorphism, in this case
269 caused by meiotic drive [1]. The measured degree of distortion at this locus (58:42) is
270 predicted by the associated decrease in pollen viability [1]. This allele is segregating in
271 about half of all individuals from a natural population [1]. Other instances of distortion loci
272 segregating at intermediate frequencies are known, but the evolutionary dynamics of
273 these cases are not as well characterized (reviewed in [5, 16])

274 A peculiarity of allelic distortion in our panel of *A. thaliana* crosses is that in most
275 cases, only a single genomic region is inherited in a non-Mendelian fashion. Classic
276 meiotic drive systems consist of a distorter locus and a responder locus, with the two
277 being almost always linked through an inversion or genetic rearrangement that reduces
278 recombination between them [5, 47-49]. As a result, classic drive loci are inherited as a
279 single distorted genomic region. Our results are reminiscent of such cases, suggesting
280 that several such loci are segregating in *A. thaliana*, although we cannot currently infer
281 the number of genes in the mapping intervals responsible for segregation distortion.

282 Apart from meiotic drive, more conventional two-locus deleterious interactions
283 conforming to the Bateson-Dobzhansky-Muller model of genetic incompatibilities can also
284 perturb expected allelic (and genotypic) segregation ratios. A survey in *D. melanogaster*
285 showed intraspecific genetic incompatibilities due to epistatic interaction between two
286 (often unlinked) loci are not uncommon, with natural strains carrying an average of 1.15
287 incompatible loci [50]. Hybrid incompatibility is a common feature in both plants in
288 animals, with many known cases of deleterious epistatic interactions between two nuclear
289 loci segregating in *A. thaliana* [51-58]. In our set of crosses, simultaneous distortion at
290 two independent genomic regions was the exception. In our design, incompatible
291 interactions would only be detectable if the F_1 was fertile and dominance relationship
292 between alleles was such that over 10% of the progeny did not give rise to seedlings. In
293 other words, if both genes acted completely recessively and the doubly homozygous
294 progeny failed to grow, they still would not be noticed in our segregation distortion scans.

295 We note that even in cases where two independent genomic regions are significantly
296 distorted in a single population, the absence of genotype data for individuals does not
297 allow us to explicitly examine if these regions genetically interact. Although the nature of
298 our experimental design has not yet revealed the species-wide architecture of partially or
299 fully recessive epistatic interactions segregating in *A. thaliana*, this can be addressed in
300 future studies by genotyping individuals instead of pools.

301 While a handful of classical segregation distortion loci has been molecularly
302 characterized in detail (reviewed in [5, 16, 26, 27]), the molecular nature of most loci is
303 still unknown. As a result, there is still much to be learned about the biological processes
304 and evolutionary forces leading to uneven segregation, including whether such alleles are
305 more likely to be evolutionarily old or young. For example, numerous cases of hybrid
306 incompatibilities in *A. thaliana* are due to interactions between disease resistance genes,
307 which have very divergent alleles, both because of rapid evolution and long-term
308 balancing selection [51, 52, 55, 56]. The fast evolution of centromeres and other satellite
309 sequence repeats, a result of intragenomic conflict, has also been shown to cause or to
310 be closely linked to allelic distortion [1, 59-61]. In our crosses, distorted regions often
311 localized near centromeres.

312 Whether the conflict arises in interspecific or intraspecific crosses, it appears that
313 natural selection, not genetic drift, is often responsible for the evolution of non-Mendelian
314 inheritance. In support of this, we found little correlation between the degree of genetic
315 differentiation between the grandparental accessions and the probability of observing
316 allelic distortion in their progeny, in line with what has been seen in a much smaller panel
317 of F₂ populations [20].

318 To conclude, by surveying a large number of F₂ populations descending from 80
319 genetically diverse grandparents, we were able to identify numerous genomic regions in
320 *A. thaliana* that are not transmitted in a Mendelian fashion. Considering that our statistical
321 power would not have allowed us to discover complete absence of genotypes resulting
322 from higher-order epistatic interactions, it is likely that the regions we identified are only
323 the tip of the iceberg. Notably, the majority of accessions tested contributed such distorted
324 alleles, emphasizing the ubiquity of alleles that are unevenly transmitted. Together, these

325 findings confirm the findings from other systems that genetic barriers segregating within
326 wild species are more common than previously thought [43, 45, 50].

327

328 **Materials and Methods**

329 *Germplasm*

330 The F₂ germplasm was generated by intercrossing 80 natural *Arabidopsis thaliana*
331 accessions for which whole-genome resequencing data had been previously published
332 [22]. Intercrossing was facilitated by induced male sterility which was achieved by artificial
333 miRNA (amiR) mediated knock-down of the floral homeotic gene *APETELA3* (*AP3*) [62].
334 One half of F₁ plants were transgene-free and able to produce self-fertilized F₂ progeny,
335 as each original female grandparent was hemizygous for the amiR transgene. In total,
336 583 F₂ populations were generated using 67 of the 80 natural accessions as the female
337 grandparent. All 80 accessions were used as the male grandparent and, on average,
338 each grandparent contributed to 14.725 F₂ populations. Germplasm information can be
339 found in Table 1.

340

341 *F₂ population growth conditions*

342 At least 300 individuals from each F₂ population were sown onto 0.5X MS media (0.7%
343 agar; pH 5.6). Prior to plating, seeds were gas sterilized for 16 hours using 40 mL of
344 household bleach (1-4%) and 1.5 mL of concentrated HCl. Seeds were stratified at 4°C
345 in the dark for 8 days and then plates were subsequently shifted to 23°C long day
346 conditions (16h light:8h dark) for 5 days. On the fifth day, all seedlings were harvested in
347 bulk from each population and flash frozen in liquid nitrogen. Plates were visibly inspected
348 to ensure high germination success (>90%).

349

350 *DNA extraction and reduced-representation library preparation*

351 DNA was extracted from each bulk using a standard CTAB preparation (2% CTAB, 1.4
352 M NaCl, 100 mM Tris (pH 8), 20 mM EDTA (pH 8)). After extraction, DNA integrity was
353 ensured by gel electrophoresis and DNA quantification was performed using the Qubit
354 fluorimeter (Qubit BR assay). For library preparation, 300 ng of each sample were diluted
355 in 27 µl. Restriction enzyme-mediated reduced-representation libraries were generated

356 using KpnI. KpnI was chosen because it cleaves the *A. thaliana* genome infrequently,
357 generating only 8,366 fragments. The library preparation protocol is detailed in [63].
358 Briefly, DNA was digested and barcoded adapter sequences with sticky ends
359 complementary to the KpnI cleavage site were ligated. After ligation, 96 barcoded F₂ bulks
360 were pooled and then sheared using the Covaris S220. Next, end-repair, dA-tailing, a
361 second universal adapter ligation, and PCR enrichment were performed using the
362 Illumina compatible NEBNext DNA Library Prep Master Mix Set. Library quality was
363 determined using the Agilent 2100 Bioanalyzer (DNA 1000 kit) and libraries were
364 normalized (10nM) based on library quantification (ng/μl) and mean fragment length.
365 Sequencing was performed on the Illumina HiSeq 2000. Adapter sequences can be found
366 in [63].

367

368 *SNP identification and allele frequency estimation*

369 The SHORE analysis program (v0.9.0) was used for all described analyses in this section
370 [64]. Sequencing reads were barcode sorted and quality filtered. During quality filtering
371 the restriction enzyme overhang was also trimmed using SHORE import. Reads for each
372 bulked population were then aligned to the TAIR10 reference genome allowing for two
373 mismatches using SHORE mapflowcell. After alignment SNPs were called with SHORE
374 qVar using the default parameters. Read counts for both the reference and non-reference
375 base were extracted for each polymorphic position. SNPs were filtered further using the
376 grandparental whole-genome information and read counts for the female grandparental
377 allele were output only for positions expected to be segregating between the two initial
378 grandparents. Based on the whole-genome sequences, a mean of 2,523 sites were
379 expected to be segregating in any F₂ population and a similar number of segregating sites
380 were observed in the reduced-representation data (Fig 1B). The allele frequency of the
381 female grandparental allele was calculated for each polymorphic position as the number
382 of reads containing the female grandparental allele divided by the total number of reads
383 covering each position.

384

385 *Modeling of allele frequency and significance testing for allelic distortion*

386 High read coverage was sought for each library to enable accurate allele frequency
387 estimation. Based on *in silico* digestions, the KpnI reduced-representation protocol was
388 expected to generate 117X coverage per F₂ library bulk. The median coverage of the
389 population bulks was 78X. The discrepancy between the expected and observed read
390 coverages was due to an excess of reads derived from the chloroplast together with lower
391 sequencer throughput. The distribution of read coverage per library is shown in Fig 1A.

392 Even with high read coverage, allele frequency estimates were still noisy. To
393 generate accurate allele frequency estimates, the allele frequency was modeled in 5 Mb
394 sliding windows (0.5 Mb step). We used a beta-binomial model to account for variation in
395 the true allele frequency as well as stochastic variation that arises from read sampling.
396 From the the optimized model we extracted the alpha and beta parameters from each
397 genomic window. These parameters describe the shape of the probability distribution in
398 each window and from these parameters the mean allele frequency as well as the 95%
399 confidence intervals were estimated. Using these estimates, a non-parametric statistical
400 test was performed to assess whether the allele frequency estimates were significantly
401 different from 50%, the expected frequency for non-distorted genomic regions. A false
402 discovery correction (FDR) was performed to account for the number of genomic windows
403 tested per population (n = 240). After allele frequency estimation, quality control measures
404 culled low quality bulks. Populations were excluded from subsequent analysis for the
405 following reasons: 1) having a genome-wide average allele frequency greater than 0.75,
406 2) exhibiting either confidence intervals (CI) larger than 0.40 or noisy confidence intervals
407 across the genome (standard deviation of CI width greater than 0.15), or 3) displaying
408 three or more chromosomes with windows that did not attain model convergence. After
409 quality control, 492 populations remained for subsequent analyses.

410

411 *Identification of distorted regions*

412 Two thresholds were used to identify significantly distorted genomic windows. The first
413 approach utilized p-value estimates from the non-parametric statistical test performed on
414 each window. False discovery rate (FDR) corrections were applied to account for the
415 number of tested genomic windows (n = 240, p < 0.05). Distorted populations were
416 required to have at least five adjacent genomic windows on the biased chromosome with

417 significant FDR corrected p-values. Populations with statistically significant segregation
418 distortion are listed in Table 1.

419 The second, less conservative, approach identified outliers by calculating Z-scores
420 for each genomic window relative to the mean allele frequency of all surveyed F_2
421 populations (0.5029). Genomic windows with allele frequency estimates greater than 2.5
422 time the population-wide standard deviation (0.0382) were considered to be distorted. A
423 distorted F_2 population was required to contain five genomic windows with significant Z-
424 scores on the chromosomes containing the locus of interest. Distorted populations
425 identified using extreme Z-scores are listed in Table 1.

426

427 *Interval identification using whole-genome resequencing*

428 Six F_2 populations were chosen for follow-up whole-genome sequencing. Selected
429 populations displayed severe distortion at one of six distinct genomic regions (Fig 9).
430 1,500 individuals were sown from each population onto 0.5X MS media (0.7% agar; pH
431 5.6) as described for the initial screen. DNA was extracted from each population bulk
432 using a standard bulk CTAB preparation (2% CTAB, 1.4 M NaCl, 100 mM Tris (pH 8), 20
433 mM EDTA (pH 8)). Illumina TruSeq libraries were prepared according to the
434 manufacturer's guidelines using 1 μ g of starting material per population. Libraries were
435 sequenced on the Illumina HiSeq 3000. K-mers (21 nt) were identified directly from the
436 short reads using jellyfish (v2.2.3) [65] with the following arguments: -m 21 -s 300M -t 10
437 -C. Not only does jellyfish identify all unique 21-mers, but it also calculates the occurrence,
438 or coverage, of each sequence. The distribution of 21-mer coverage is shown in Figure 8
439 for each population. Their coverage distribution is bimodal. This first peak (~25X)
440 represents sequences found in only one of the two grandparents' genomes, while the
441 second peak represents those sequences found in both grandparental genomes. 21-mers
442 found in only one of the two grandparental genomes (coverage < 25X) were aligned to
443 the TAIR10 genome using bwa aln. Only perfect matches were allowed. A 1 Mb sliding
444 window (50 kb step) was used to plot the 21-mer coverage across the distorted
445 chromosome in each population. Regions of the genome with reduced coverage of 21-
446 mers are located within the candidate interval (Fig 9). Interval boundaries were delineated
447 by merging all windows within 1X coverage of the minimal window in the candidate region.

448

449 *Interval identification for distortion bulk segregant analysis (BSA)*

450 Bulk segregant analysis [25] was used to narrow the candidate intervals for 3 genomic
451 regions. Alleles from three grandparental accessions (Star-8, ICE49, ICE63) were
452 repeatedly favored in over 40% of their respective F₂ crosses (based on the Z-score
453 outlier threshold). Sequencing reads from the original screen were combined for all
454 distorted populations sharing the grandparent of interest into a distorted bulk. Those that
455 shared the grandparent, but did not exhibit distortion were combined into a normal bulk.
456 Positions segregating between the grandparent of interest and all other members of the
457 bulk were identified. The positions segregating in the distorted bulk are not shared with
458 those segregating in the normal bulk. By combining reads from multiple populations, a
459 median of 806 to 1135X coverage was achieved at each segregating position. Candidate
460 intervals were calculated from the maximally distorted position to any flanking segregating
461 site that was within 5% of the peak allele frequency (Table 2).

462 **References**

- 463 1. Fishman L, Saunders A (2008) Centromere-associated female meiotic drive entails
464 male fitness costs in monkeyflowers. *Science* 322(5907):1559-62.
- 465 2. Phadnis N, Orr HA (2009) A single gene causes both male sterility and segregation
466 distortion in *Drosophila* hybrids. *Science* 323(5912):376-9.
- 467 3. McDermott SR, Noor MA (2010) The role of meiotic drive in hybrid male sterility. *Philos*
468 *Trans R Soc Lond B Biol Sci* 365(1544):1265-72.
- 469 4. Presgraves DC (2010) The molecular evolutionary basis of species formation. *Nat Rev*
470 *Genet* 11(3):175-80.
- 471 5. Lyttle TW (1991) Segregation distorters. *Annu Rev Genet* 25:511-57.
- 472 6. Orr HA (1996) Dobzhansky, Bateson, and the genetics of speciation. *Genetics*
473 144(4):1331-5.
- 474 7. Orr HA, Presgraves DC (2000) Speciation by postzygotic isolation: forces, genes and
475 molecules. *Bioessays* 22(12):1085-94.
- 476 8. Bomblies K, Weigel D (2007) Hybrid necrosis: autoimmunity as a potential gene-flow
477 barrier in plant species. *Nat Rev Genet* 8(5):382-93.
- 478 9. Sandler L, Hiraizumi Y, Sandler I (1959) Meiotic drive in natural populations of
479 *Drosophila melanogaster*. I. the cytogenetic basis of segregation-distortion.
480 *Genetics* 44(2):233-50.
- 481 10. Malik HS, Henikoff S (2002) Conflict begets complexity: the evolution of centromeres.
482 *Curr Opin Genet Dev* 12(6):711-8.
- 483 11. Sturtevant AH, Dobzhansky T (1936) Geographical distribution and cytology of "sex
484 ratio" in *drosophila pseudoobscura* and related species. *Genetics* 21(4):473-90.
- 485 12. Hartl DL, Hiraizum.Y, Crow JF (1967) Evidence for sperm dysfunction as mechanism
486 of segregation distortion in *Drosophila melanogaster*. *Proc Natl Acad Sci U S A*
487 58(6):2240-5.
- 488 13. Rhoades MM (1942) Preferential segregation in maize. *Genetics* 27(4):0395-407.
- 489 14. Rhoades MM, Dempsey E, Ghidoni A (1967) Chromosome elimination in maize
490 induced by supernumerary B chromosomes. *Proc Natl Acad Sci U S A* 57(6):1626-
491 32.
- 492 15. Dunn LC, Bennett D (1968) A new case of transmission ratio distortion in house
493 mouse. *Proc Natl Acad Sci U S A* 61(2):570-3.
- 494 16. Zimmering S, Sandler L, Nicolett B (1970) Mechanisms of meiotic drive. *Annu Rev*
495 *Genet* 4:409-36.
- 496 17. Snow AA, Spira TP, Liu H (2000) Effects of sequential pollination on the success of
497 "fast" and "slow" pollen donors in *Hibiscus moscheutos* (Malvaceae). *Am J Bot*
498 87(11):1656-9.
- 499 18. Lister C, Dean C (1993) Recombinant inbred lines for mapping RFLP and phenotypic
500 markers in *Arabidopsis thaliana*. *Plant J* 4(4):745-50.
- 501 19. Alonso-Blanco C, Peeters AJ, Koornneef M, Lister C, Dean C, van den Bosch N, et
502 al. (1998) Development of an AFLP based linkage map of Ler, Col and Cvi
503 *Arabidopsis thaliana* ecotypes and construction of a Ler/Cvi recombinant inbred
504 line population. *Plant J* 14(2):259-71.
- 505 20. Salomé PA, Bomblies K, Fitz J, Laitinen RA, Warthmann N, Yant L, et al. (2012) The
506 recombination landscape in *Arabidopsis thaliana* F₂ populations. *Heredity*
507 108(4):447-55.

- 508 21. Bomblies K, Yant L, Laitinen R, Kim S-T, Hollister JD, Warthmann N, et al. (2010)
509 Local-scale patterns of genetic variability, outcrossing and spatial structure in
510 natural stands of *Arabidopsis thaliana*. PLoS Genet 6(3):e1000890.
- 511 22. Cao J, Schneeberger K, Ossowski S, Günther T, Bender S, Fitz J, et al. (2011) Whole-
512 genome sequencing of multiple *Arabidopsis thaliana* populations. Nat Genet
513 43:956-63.
- 514 23. Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, et al. (2008) Rapid
515 SNP discovery and genetic mapping using sequenced RAD markers. PLoS ONE
516 3(10):e3376.
- 517 24. Monson-Miller J, Sanchez-Mendez DC, Fass J, Henry IM, Tai TH, Comai L (2012)
518 Reference genome-independent assessment of mutation density using restriction
519 enzyme-phased sequencing. BMC Genomics 13:72.
- 520 25. Michelmore RW, Paran I, Kesseli RV (1991) Identification of markers linked to
521 disease-resistance genes by bulked segregant analysis: a rapid method to detect
522 markers in specific genomic regions by using segregating populations. Proc Natl
523 Acad Sci U S A 88(21):9828-32.
- 524 26. Lyon MF (2003) Transmission ratio distortion in mice. Annu Rev Genet 37:393-408.
- 525 27. Larracuenta AM, Presgraves DC (2012) The selfish Segregation Distorter gene
526 complex of *Drosophila melanogaster*. Genetics 192(1):33-53.
- 527 28. Hammond TM, Rehard DG, Xiao H, Shiu PK (2012) Molecular dissection of
528 *Neurospora* Spore killer meiotic drive elements. Proc Natl Acad Sci U S A
529 109(30):12093-8.
- 530 29. Cameron DR, Moav RM (1957) Inheritance in *Nicotiana tabacum* Xxvii. Pollen Killer,
531 an alien genetic locus inducing abortion of microspores not carrying it. Genetics
532 42(3):326-35.
- 533 30. Maguire MP (1963) High transmission frequency of a *Tripsacum* chromosome in corn.
534 Genetics 48(9):1185-94.
- 535 31. Siracusa LD, Alvord WG, Bickmore WA, Jenkins NA, Copeland NG (1991)
536 Interspecific backcross mice show sex-specific differences in allelic inheritance.
537 Genetics 128(4):813-21.
- 538 32. Tao Y, Hartl DL, Laurie CC (2001) Sex-ratio segregation distortion associated with
539 reproductive isolation in *Drosophila*. Proc Natl Acad Sci U S A 98(23):13183-8.
- 540 33. Zanders SE, Eickbush MT, Yu JS, Kang JW, Fowler KR, Smith GR, et al. (2014)
541 Genome rearrangements and pervasive meiotic drive cause hybrid infertility in
542 fission yeast. Elife 3:e02630.
- 543 34. Belanger S, Esteves P, Clermont I, Jean M, Belzile F (2016) Genotyping-by-
544 sequencing on pooled samples and its use in measuring segregation bias during
545 the course of androgenesis in barley. Plant Genome 9(1).
- 546 35. Cui Y, Zhang F, Xu J, Li Z, Xu S (2015) Mapping quantitative trait loci in selected
547 breeding populations: A segregation distortion approach. Heredity (Edinb)
548 115(6):538-46.
- 549 36. Belanger S, Clermont I, Esteves P, Belzile F (2016) Extent and overlap of segregation
550 distortion regions in 12 barley crosses determined via a Pool-GBS approach.
551 Theor Appl Genet 129(7):1393-404.
- 552 37. Wei KH, Reddy HM, Rathnam C, Lee J, Lin D, Ji S, et al. (2017) A pooled sequencing
553 approach identifies a candidate meiotic driver in *Drosophila*. Genetics.

- 554 38. Hiraizumi Y, Thomas AM (1984) Suppressor systems of Segregation Distorter (SD)
555 chromosomes in natural populations of DROSOPHILA MELANOGASTER.
556 Genetics 106(2):279-92.
- 557 39. Hammer MF, Schimenti J, Silver LM (1989) Evolution of mouse chromosome 17 and
558 the origin of inversions associated with t haplotypes. Proc Natl Acad Sci U S A
559 86(9):3261-5.
- 560 40. Hickey WA, Craig GB, Jr. (1966) Distortion of sex ratio in populations of Aedes
561 aegypti. Can J Genet Cytol 8(2):260-78.
- 562 41. Perkins DD, Barry EG (1977) The cytogenetics of Neurospora. Adv Genet 19:133-
563 285.
- 564 42. McMullen MD, Kresovich S, Villeda HS, Bradbury P, Li H, Sun Q, et al. (2009) Genetic
565 properties of the maize nested association mapping population. Science
566 325(5941):737-40.
- 567 43. Hou J, Friedrich A, Gounot JS, Schacherer J (2015) Comprehensive survey of
568 condition-specific reproductive isolation reveals genetic incompatibility in yeast.
569 Nat Commun 6:7214.
- 570 44. Fragoso CA, Moreno M, Wang Z, Heffelfinger C, Arbelaez LJ, Aguirre JA, et al. (2017)
571 Genetic architecture of a rice nested association mapping population. G3
572 7(6):1913-26.
- 573 45. Seidel HS, Rockman MV, Kruglyak L (2008) Widespread genetic incompatibility in *C.*
574 *elegans* maintained by balancing selection. Science 319(5863):589-94.
- 575 46. Ben-David E, Burga A, Kruglyak L (2017) A maternal-effect selfish genetic element in
576 *Caenorhabditis elegans*. Science 356(6342):1051-5.
- 577 47. Silver LM (1985) Mouse t haplotypes. Annu Rev Genet 19:179-208.
- 578 48. Stalker HD (1961) The genetic systems modifying meiotic drive in *Drosophila*
579 *paramelanica*. Genetics 46(2):177-202.
- 580 49. Wu CI, Beckenbach AT (1983) Evidence for extensive genetic differentiation between
581 the sex-ratio and the standard arrangement of DROSOPHILA
582 PSEUDOOBSCURA and *D. PERSIMILIS* and identification of hybrid sterility
583 factors. Genetics 105(1):71-86.
- 584 50. Corbett-Detig RB, Zhou J, Clark AG, Hartl DL, Ayroles JF (2013) Genetic
585 incompatibilities are widespread within species. Nature 504(7478):135-7.
- 586 51. Bomblies K, Lempe J, Epple P, Warthmann N, Lanz C, Dangl JL, et al. (2007)
587 Autoimmune response as a mechanism for a Dobzhansky-Muller-type
588 incompatibility syndrome in plants. PLoS Biol 5(9):e236.
- 589 52. Alcázar R, Garcia AV, Parker JE, Reymond M (2009) Incremental steps toward
590 incompatibility revealed by *Arabidopsis* epistatic interactions modulating salicylic
591 acid pathway activation. Proc Natl Acad Sci USA 106(1):334-9.
- 592 53. Bikard D, Patel D, Le Mette C, Giorgi V, Camilleri C, Bennett MJ, et al. (2009)
593 Divergent evolution of duplicate genes leads to genetic incompatibilities within *A.*
594 *thaliana*. Science 323(5914):623-6.
- 595 54. Vlad D, Rappaport F, Simon M, Loudet O (2010) Gene transposition causing natural
596 variation for growth in *Arabidopsis thaliana*. PLoS Genet 6(5):e1000945.
- 597 55. Durand S, Bouche N, Perez Strand E, Loudet O, Camilleri C (2012) Rapid
598 establishment of genetic incompatibility through natural epigenetic variation. Curr
599 Biol 22(4):326-31.

- 600 56. Chae E, Bomblies K, Kim ST, Karelina D, Zaidem M, Ossowski S, et al. (2014)
601 Species-wide genetic incompatibility analysis identifies immune genes as hot
602 spots of deleterious epistasis. *Cell* 159(6):1341-51.
- 603 57. Agorio A, Durand S, Fiume E, Brousse C, Gy I, Simon M, et al. (2017) An *Arabidopsis*
604 natural epiallele maintained by a feed-forward silencing loop between histone and
605 DNA. *PLoS Genet* 13(1):e1006551.
- 606 58. Plötner B, Nurmi M, Fischer A, Watanabe M, Schneeberger K, Holm S, et al. (2017)
607 Chlorosis caused by two recessively interacting genes reveals a role of RNA
608 helicase in hybrid breakdown in *Arabidopsis thaliana*. *Plant J*.
- 609 59. Wu CI, Lyttle TW, Wu ML, Lin GF (1988) Association between a satellite DNA
610 sequence and the Responder of Segregation Distorter in *D. melanogaster*. *Cell*
611 54(2):179-89.
- 612 60. Chmatal L, Gabriel SI, Mitsainas GP, Martinez-Vargas J, Ventura J, Searle JB, et al.
613 (2014) Centromere strength provides the cell biological basis for meiotic drive and
614 karyotype evolution in mice. *Curr Biol* 24(19):2295-300.
- 615 61. Maheshwari S, Tan EH, West A, Franklin FC, Comai L, Chan SW (2015) Naturally
616 occurring differences in CENH3 affect chromosome segregation in zygotic mitosis
617 of hybrids. *PLoS Genet* 11(1):e1004970.
- 618 62. Chae E, Bomblies K, Kim ST, Karelina D, Zaidem M, Ossowski S, et al. (2014)
619 Species-wide genetic incompatibility analysis identifies immune genes as hot
620 spots of deleterious epistasis. *Cell* 159(6):1341-51.
- 621 63. Rowan BA, Seymour DK, Chae E, Lundberg DS, Weigel D (2017) Methods for
622 genotyping-by-sequencing. *Methods in molecular biology* 1492:221-42.
- 623 64. Ossowski S, Schneeberger K, Clark RM, Lanz C, Warthmann N, Weigel D (2008)
624 Sequencing of natural strains of *Arabidopsis thaliana* with short reads. *Genome*
625 *Res* 18:2024-33.
- 626 65. Marcais G, Kingsford C (2011) A fast, lock-free approach for efficient parallel counting
627 of occurrences of k-mers. *Bioinformatics* 27(6):764-70.
- 628

629 **Figure legends**

630 Figure 1. Reduced-representation sequencing reliably enriches for 1% of the *A. thaliana*
631 genome. A) Distribution of the mean sequencing coverage at sites segregating in each
632 F₂ population. B) Distribution of the number of sites segregating in each F₂ population.
633 The mean observed number of segregating sites (2,500) is comparable to the expected
634 number of segregating sites derived from previously published resequencing data [22].

635
636 Figure 2. Statistically significant segregation distortion is evident across a wide range of
637 germplasm combinations. Genotypic combinations surveyed in this F₂ screen are shown
638 in blue. Combinations not surveyed are indicated in black. Populations exhibiting
639 significant segregation distortion based on non-parametric statistical tests of beta-
640 binomial modeled allele frequencies are shown in green. Grandparental accessions are
641 ordered by the geographic location of their collection [22]. Female grandparents are
642 located on the y-axis and male grandparents on the x-axis.

643
644 Figure 3. Z-score estimated segregation distortion is evident across a wide range of
645 germplasm combinations. Genotypic combinations surveyed in this F₂ screen are shown
646 in blue. Combinations not surveyed are indicated in black. Populations exhibiting
647 significant segregation distortion based on Z-score metrics are shown in green.
648 Grandparental accessions are ordered by the geographic location of their collection [22].
649 Female grandparents are located on the y-axis and male grandparents on the x-axis.

650
651 Figure 4. Example of a representative F₂ population exhibiting significant segregation
652 distortion. Distortion in this population was detected by two methods: 1) by significant
653 statistical tests of deviation of the beta-binomial modeled allele frequencies from the
654 expected frequency of 0.5 and 2) by significant Z-score deviation (2.5X) from the
655 population-wide allele frequency mean (0.5029). The beta-binomial modeled allele
656 frequency (green) across each chromosome is plotted in the upper panel. 95%
657 confidence intervals are indicated by the shaded grey area and the expected frequency
658 of 0.5 is marked by the dashed black line. The lower panel plots the $-\log_{10}$ (p-value)

659 derived from the non-parametric statistical test. The dashed black line in this panel
660 represents the FDR corrected ($n = 240$) significance threshold ($p < 0.05$).

661

662 Figure 5. Genomic properties of distorted loci. A) The fraction of surveyed F_2 populations
663 that exhibit segregation distortion at either one or two genomic loci. B) The number of
664 populations containing distorted loci that reside on each of the five *A. thaliana*
665 chromosomes.

666

667 Figure 6. Many grandparental accessions contribute biased alleles. Each grandparent
668 contributes its genetic material to a median of 14 distinct F_2 populations. Plotted are the
669 percent of F_2 populations sharing a grandparent that are significantly distorted as
670 measured either by A) Z-score metrics, or B) FDR corrected statistical tests based on
671 beta-binomial modeled allele frequencies. Whether the grandparent was used as the
672 male (light color) or female (dark color) is also indicated.

673

674 Figure 7. Genetic distance between grandparental accessions is not predictive of biased
675 allelic transmission. A box plot of the genetic distance between the grandparental
676 accessions of normal (grey) and distorted (colored) F_2 populations. At a significance
677 threshold of $p < 0.01$, the genetic distance between grandparents of distorted populations
678 determined from outlier (purple) or FDR (green) approaches is not significantly different
679 from that of normal populations (Wilcoxon rank sum test). Genetic distance was
680 calculated as the number of segregating sites over the number of interrogated sites. All
681 positions were required to have complete coverage across all 80 grandparental
682 accessions.

683

684 Figure 8. Distribution of whole-genome resequencing coverage of unique 21-mers. The
685 coverage of unique 21 nt k-mers is plotted for each of the six populations that underwent
686 whole-genome resequencing. The first peak in coverage represents 21-mers found in
687 only one of the two grandparents, while the second, larger peak represents those
688 sequences found in both.

689

690 Figure 9. Candidate intervals refined using 21-mer coverage. For each population, the
691 upper panel displays the beta-binomial modeled allele frequency estimates (green) and
692 their 95% confidence intervals (grey) as described in Figure 4. In the lower panel, the
693 coverage of 21-mers unique to only one of the two grandparents (coverage < 25X) is
694 plotted in 1 Mb sliding windows (50 kb step). Coverage decreases in the candidate
695 regions. Intervals (grey box) are defined by merging windows within 1X coverage of the
696 minimal window in each population. No candidate region was defined for POP064 as the
697 coverage decrease coincides with the centromere, not the distorted region.

698

699 Figure 10. Increasing segregants narrows candidate intervals. Bulked segregant analysis
700 was performed for grandparental accessions that repeatedly contributed distorted loci
701 (Star-8 (A), ICE63 (B), and ICE49). Sequencing reads were combined for populations
702 exhibiting distortion or not when crossed to the focal grandparent. An average of over
703 800X coverage was achieved at sites segregating between the focal accessions and all
704 other members in the bulk. Candidate intervals (grey box) merged all segregating
705 positions within 5% of the maximal marker's allele frequency. Data for ICE49 are not
706 shown as there were too few segregating sites.

707

708 **Table legends**

709 Table 1. Germplasm information for surveyed F₂ populations. The grandparents of each
710 F₂ population are listed as well as whether the population had sufficient data quality for
711 subsequent analysis. Populations with significant distortion are indicated for both
712 thresholds (FDR and outlier).

713

714 Table 2. Candidate intervals for distorted loci. Candidate intervals identified through both
715 k-mer and bulk segregant analysis are listed.

Figure 1

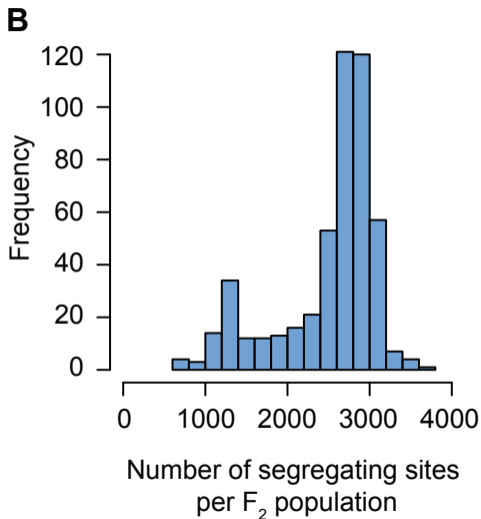
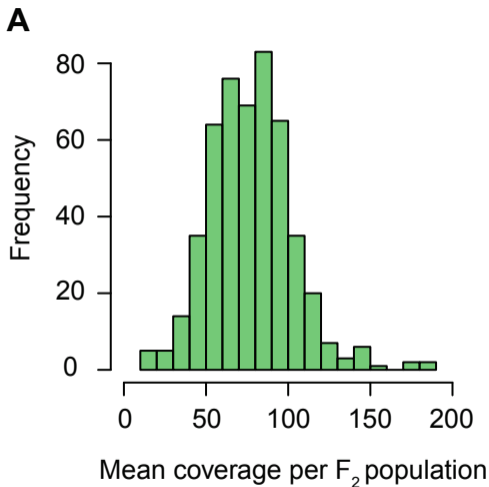
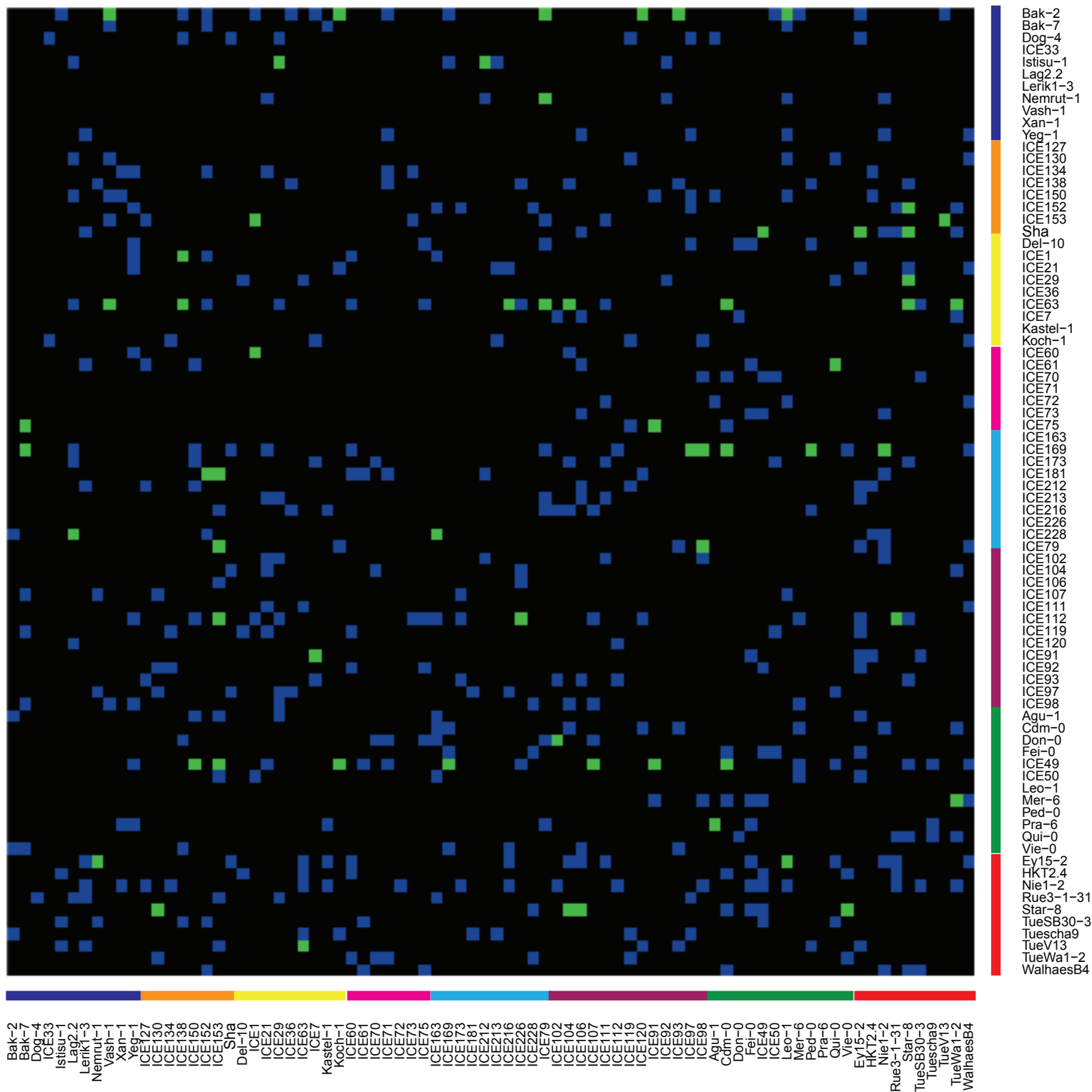


Figure 2



- Caucasus
- Central Asia
- Eastern Europe
- Russia
- South Tyrol
- Southern Italy
- Spain / North Africa
- Swabia

Figure 3

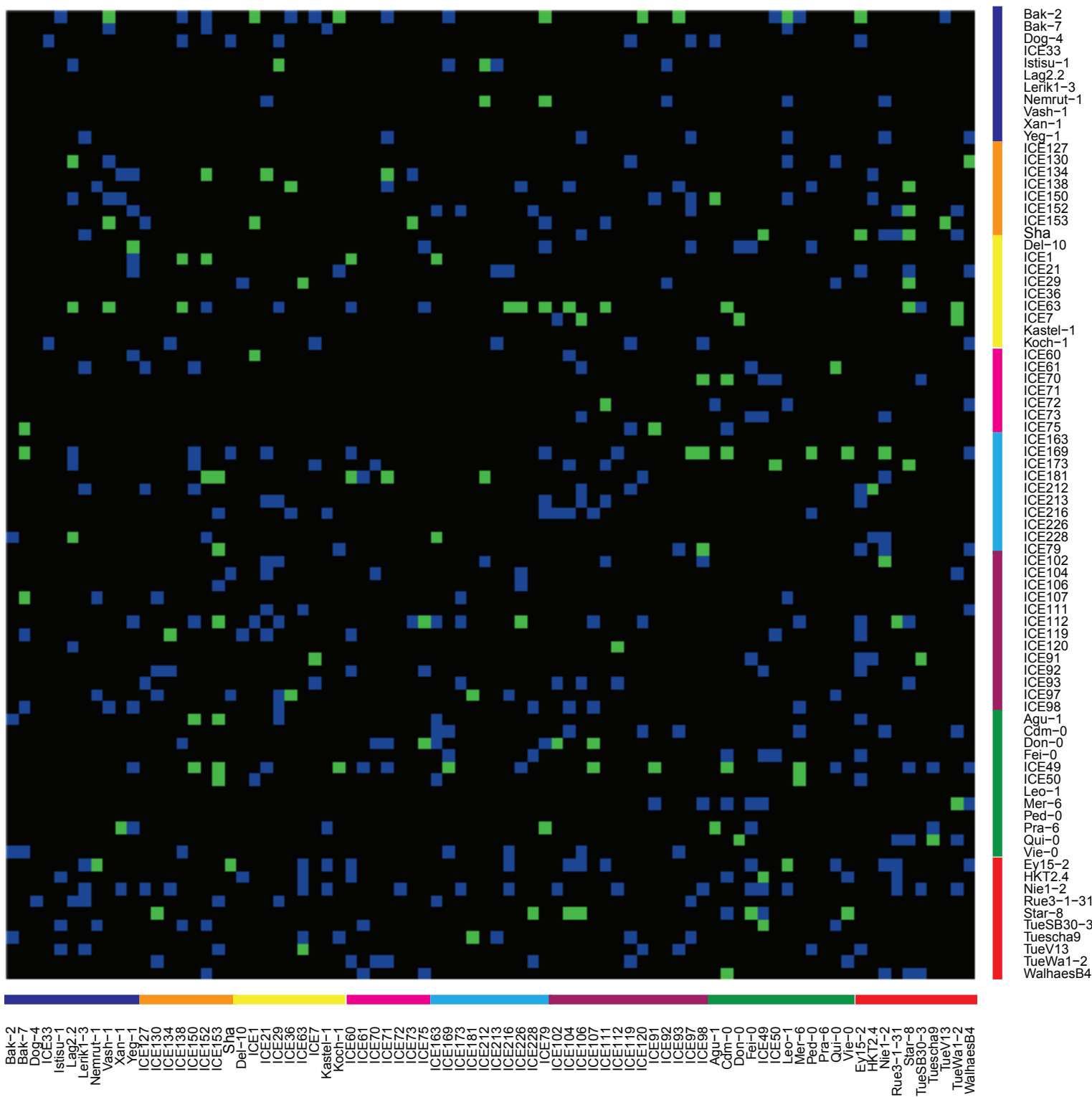


Figure 4

POP035: ICE63 x Vash-1

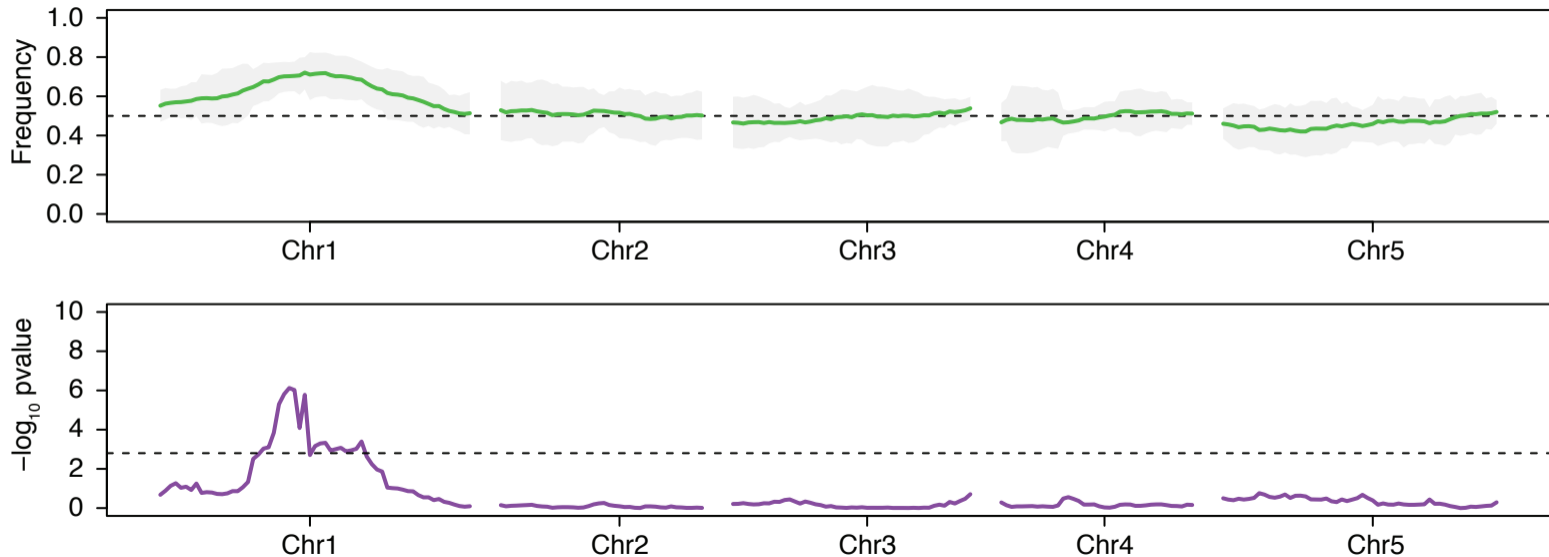


Figure 5

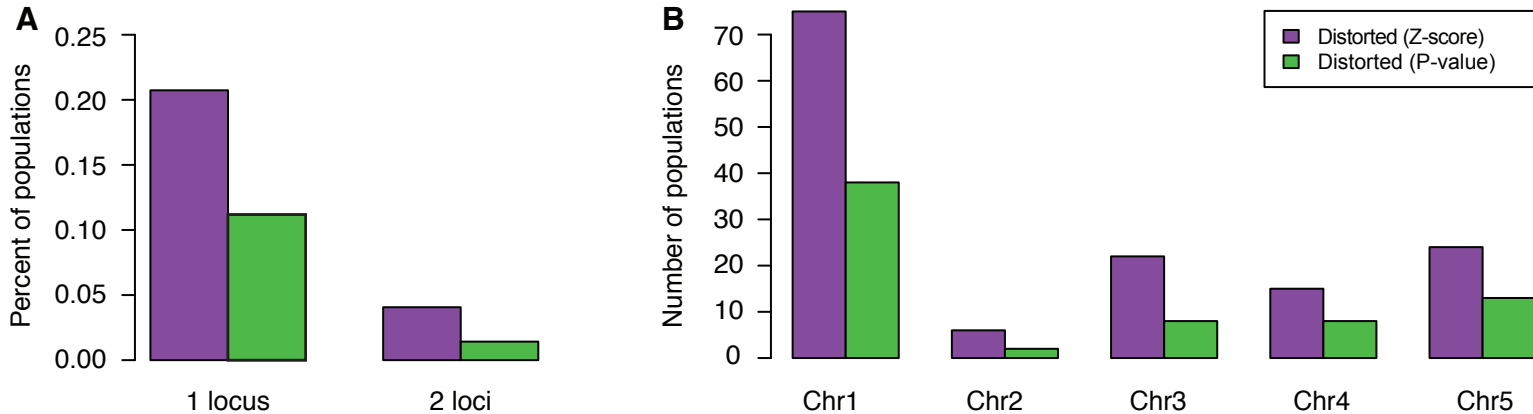


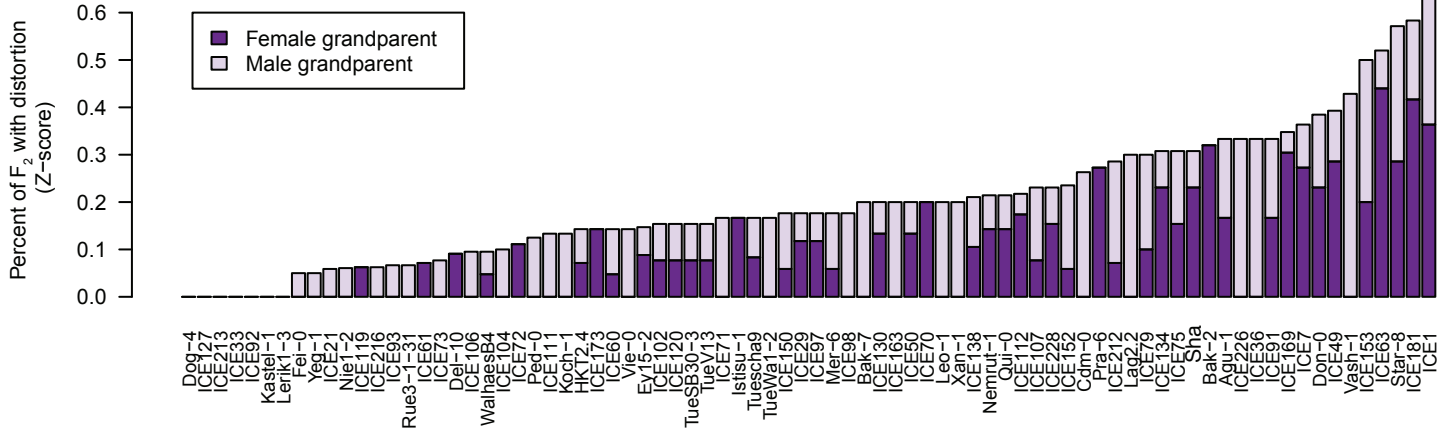
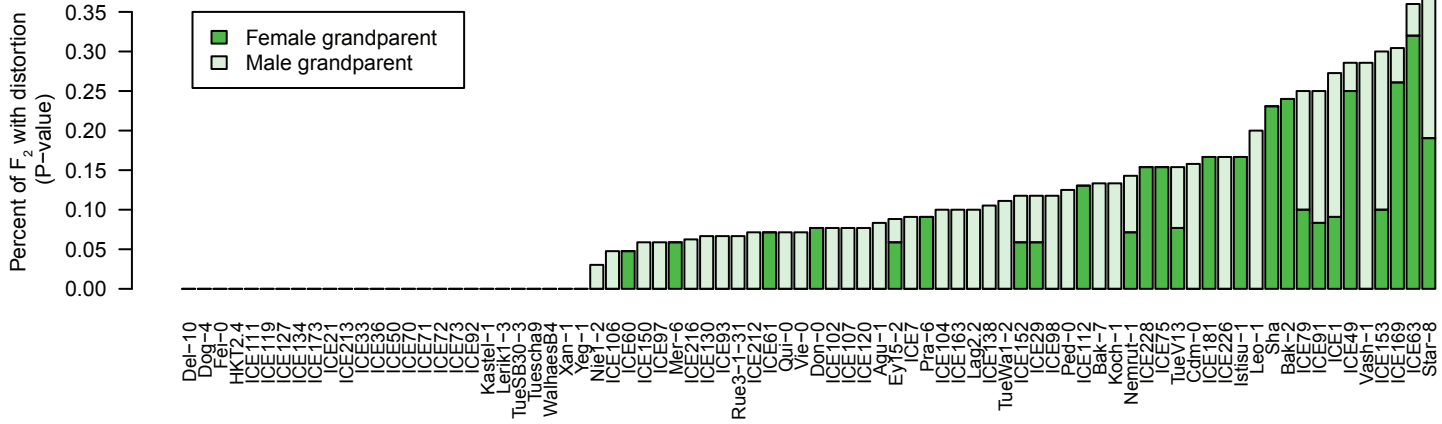
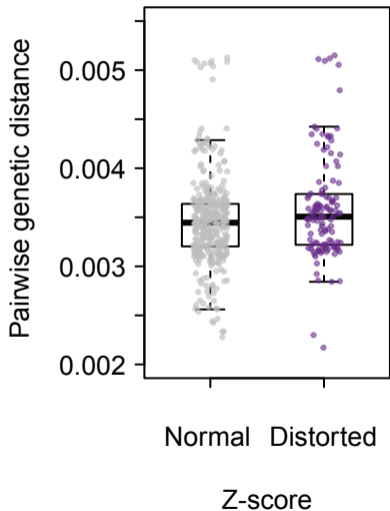
Figure 6**A****B**

Figure 7

A



B

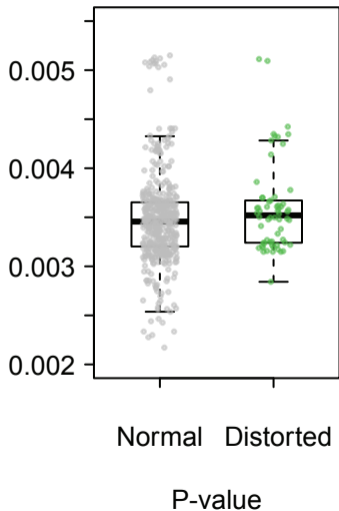
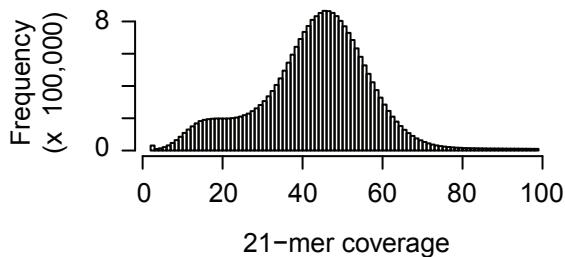
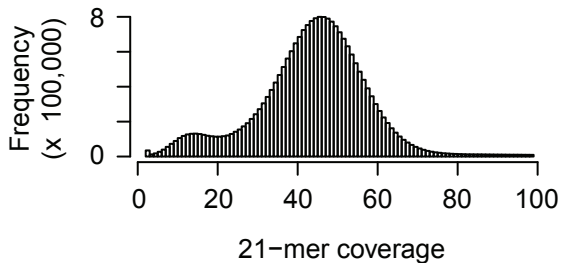


Figure 8

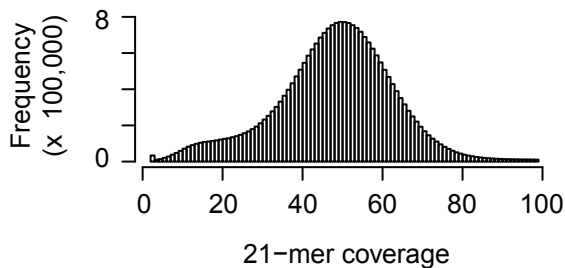
POP007



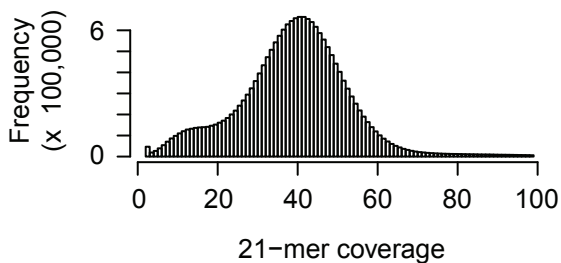
POP026



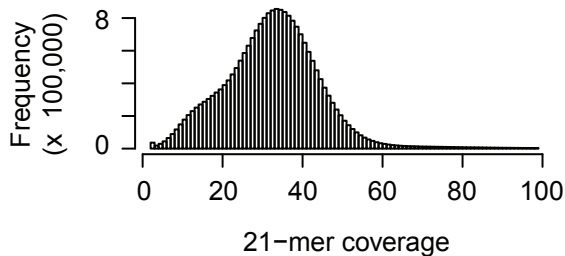
POP035



POP063



POP064



POP100

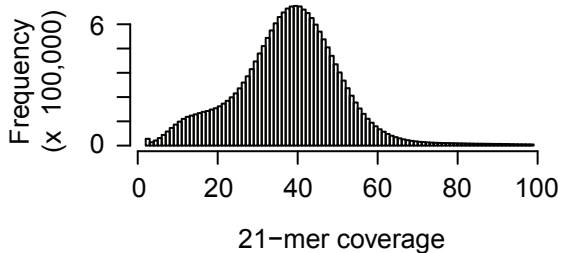
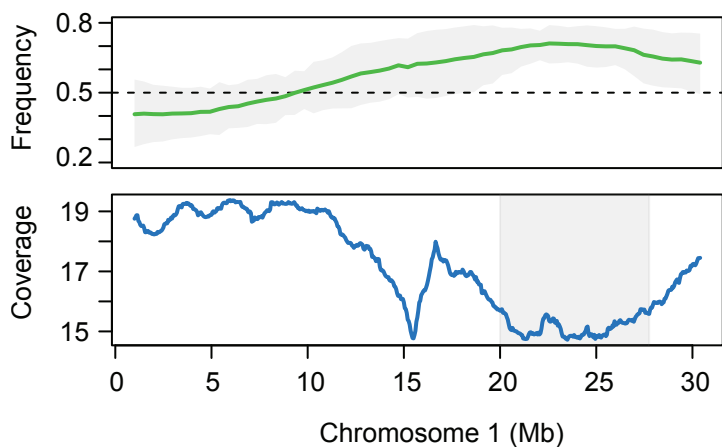
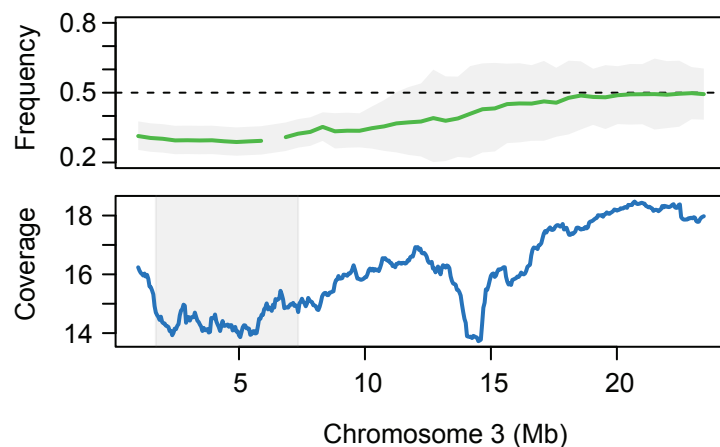


Figure 9

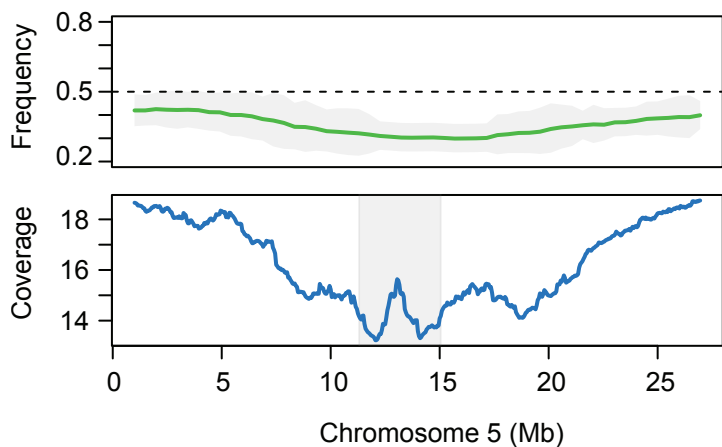
POP007: ICE49 x ICE153



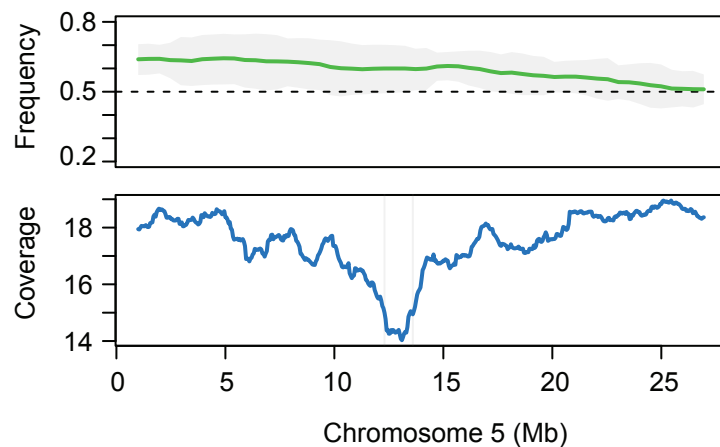
POP063: ICE169 x Bak-7



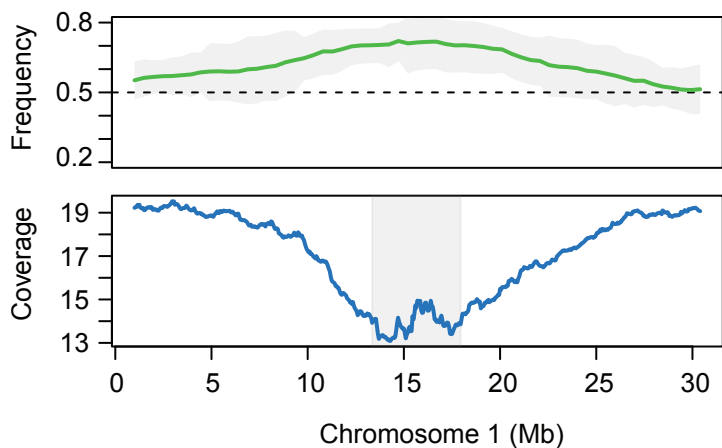
POP026: ICE63 x ICE216



POP064: ICE169 x Cdm-0



POP035: ICE63 x Vash-1



POP100: Ey15.2 x Leo-1

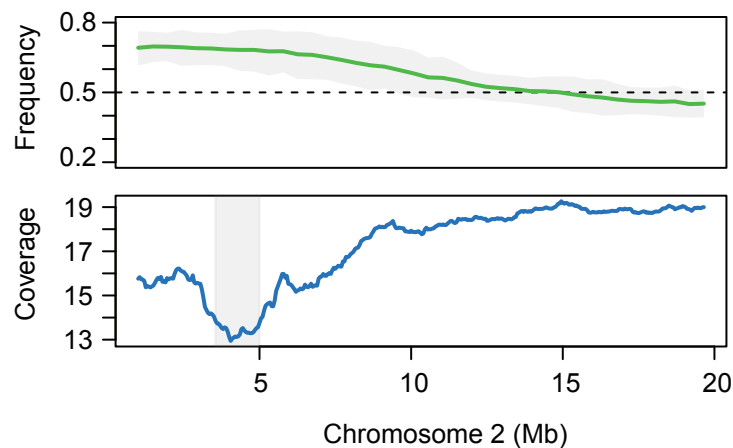


Figure 10

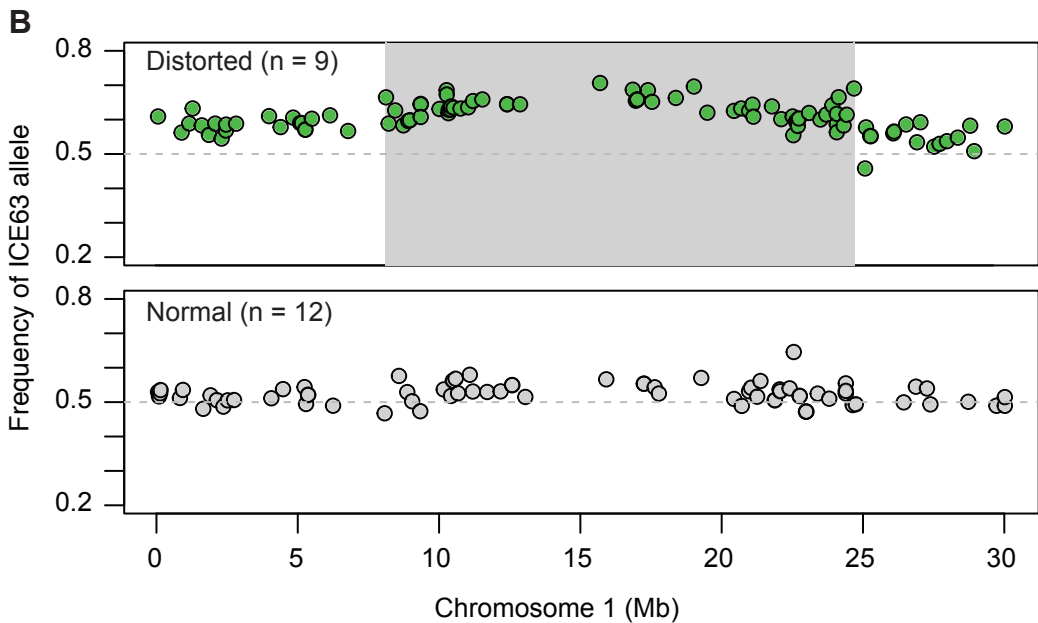
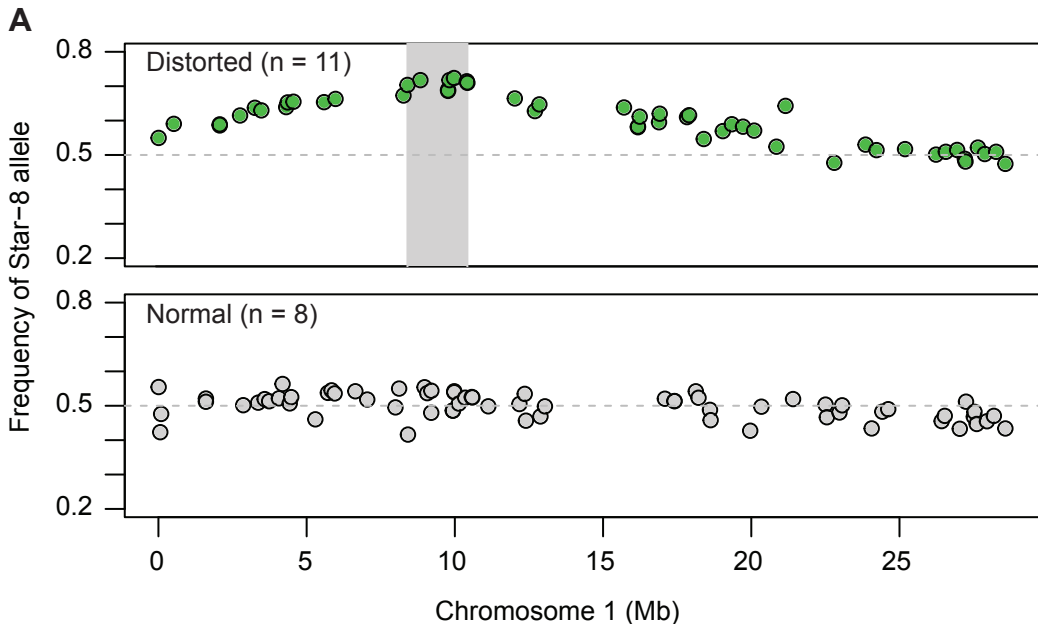


Table 1. Germplasm information for surveyed F₂ populations.

Population ID	Female grandparent	Male grandparent	Passed QC (1=yes, 0=no)	Distorted (P-value) (1=yes, 0=no)	Distorted (Z-score) (1=yes, 0=no)	Follow-up sequencing (1=yes, 0=no)
POP001	ICE49	ICE21	1	0	0	0
POP002	ICE49	ICE61	1	0	0	0
POP003	ICE49	ICE71	1	0	0	0
POP004	ICE49	ICE91	1	1	1	0
POP005	ICE49	ICE107	1	1	1	0
POP006	ICE49	ICE150	1	1	1	0
POP007	ICE49	ICE153	1	1	1	1
POP008	ICE49	ICE169	1	1	1	0
POP009	ICE49	ICE226	1	0	0	0
POP010	ICE49	Cdm-0	1	1	1	0
POP011	ICE49	Koch-1	1	1	1	0
POP012	ICE49	Mer-6	1	0	1	0
POP013	ICE49	Qui-0	1	0	0	0
POP014	ICE49	Star-8	1	0	0	0
POP015	ICE49	Tuescha9	1	0	0	0
POP016	ICE49	WalhaesB4	1	0	0	0
POP017	ICE49	Yeg-1	1	0	0	0
POP018	ICE63	ICE29	1	0	0	0
POP019	ICE63	ICE60	1	0	0	0
POP020	ICE63	ICE75	1	0	0	0
POP021	ICE63	ICE79	1	1	1	0
POP022	ICE63	ICE104	1	1	1	0
POP023	ICE63	ICE111	1	0	1	0
POP024	ICE63	ICE138	1	1	1	0
POP025	ICE63	ICE152	1	0	0	0
POP026	ICE63	ICE216	1	1	1	1
POP027	ICE63	ICE226	1	0	1	0
POP028	ICE63	Cdm-0	1	1	1	0
POP029	ICE63	Don-0	0	0	0	0
POP030	ICE63	Lag2.2	1	0	1	0
POP031	ICE63	Pra-6	0	0	0	0
POP032	ICE63	Star-8	1	1	1	0
POP033	ICE63	TueSB30-3	1	0	0	0
POP034	ICE63	TueWa1-2	1	1	1	0
POP035	ICE63	Vash-1	1	1	1	1
POP036	ICE112	ICE1	1	0	0	0
POP037	ICE112	ICE29	1	0	0	0
POP038	ICE112	ICE73	1	0	0	0
POP039	ICE112	ICE75	1	0	1	0
POP040	ICE112	ICE107	1	0	0	0

POP041	ICE112	ICE120	1	0	0	0
POP042	ICE112	ICE150	1	0	0	0
POP043	ICE112	ICE153	1	1	1	0
POP044	ICE112	ICE163	1	0	0	0
POP045	ICE112	ICE173	1	0	0	0
POP046	ICE112	ICE226	1	1	1	0
POP047	ICE112	Don-0	0	0	0	0
POP048	ICE112	Ey15-2	1	0	0	0
POP049	ICE112	Mer-6	1	0	0	0
POP050	ICE112	Rus3-1-31	1	1	1	0
POP051	ICE112	Star-8	1	0	0	0
POP052	ICE112	Tuescha9	0	0	0	0
POP053	ICE112	Yeg-1	1	0	0	0
POP054	ICE169	ICE21	1	0	0	0
POP055	ICE169	ICE60	1	0	0	0
POP056	ICE169	ICE61	0	0	0	0
POP057	ICE169	ICE79	1	0	0	0
POP058	ICE169	ICE97	1	1	1	0
POP059	ICE169	ICE98	1	1	1	0
POP060	ICE169	ICE112	1	0	0	0
POP061	ICE169	ICE150	1	0	0	0
POP062	ICE169	ICE173	0	0	0	0
POP063	ICE169	Bak-7	1	1	1	0
POP064	ICE169	Cdm-0	1	1	1	1
POP065	ICE169	Lag2.2	1	0	0	0
POP066	ICE169	Nie1-2	1	1	1	0
POP067	ICE169	Ped-0	1	1	1	0
POP068	ICE169	Sha	1	0	0	0
POP069	ICE169	Vie-0	1	0	1	0
POP070	ICE169	WalhaesB4	1	0	0	0
POP071	Bak-2	ICE1	1	0	1	0
POP072	Bak-2	ICE7	1	0	0	0
POP073	Bak-2	ICE36	1	0	0	0
POP074	Bak-2	ICE50	1	0	0	0
POP075	Bak-2	ICE71	1	0	0	0
POP076	Bak-2	ICE79	1	1	1	0
POP077	Bak-2	ICE93	1	1	1	0
POP078	Bak-2	ICE120	1	1	1	0
POP079	Bak-2	ICE138	1	0	0	0
POP080	Bak-2	ICE152	1	0	0	0
POP081	Bak-2	ICE169	1	0	0	0
POP082	Bak-2	Ey15-2	1	0	1	0

POP083	Bak-2	Istisu-1	1	0	0	0
POP084	Bak-2	Koch-1	1	1	1	0
POP085	Bak-2	Leo-1	1	1	1	0
POP086	Bak-2	Mer-6	1	0	0	0
POP087	Bak-2	TueV13	1	0	0	0
POP088	Bak-2	Vash-1	1	1	1	0
POP089	Ey15-2	ICE60	1	0	0	0
POP090	Ey15-2	ICE61	0	0	0	0
POP091	Ey15-2	ICE63	1	0	0	0
POP092	Ey15-2	ICE104	1	0	0	0
POP093	Ey15-2	ICE106	1	0	0	0
POP094	Ey15-2	ICE111	1	0	0	0
POP095	Ey15-2	ICE127	0	0	0	0
POP096	Ey15-2	ICE181	0	0	0	0
POP097	Ey15-2	ICE216	1	0	0	0
POP098	Ey15-2	Fei-0	1	0	0	0
POP099	Ey15-2	Kastel-1	1	0	0	0
POP100	Ey15-2	Leo-1	1	1	1	1
POP101	Ey15-2	Lerik1-3	1	0	0	0
POP102	Ey15-2	Nemrut-1	1	1	1	0
POP103	Ey15-2	Nie1-2	1	0	0	0
POP104	Ey15-2	Qui-0	1	0	0	0
POP105	Ey15-2	Rus3-1-31	1	0	0	0
POP106	Ey15-2	Sha	1	0	1	0
POP107	Ey15-2	WalhaesB4	1	0	0	0
POP108	Nie1-2	ICE49	1	0	0	0
POP109	Nie1-2	ICE63	1	0	0	0
POP110	Nie1-2	ICE72	1	0	0	0
POP111	Nie1-2	ICE98	1	0	0	0
POP112	Nie1-2	ICE102	1	0	0	0
POP113	Nie1-2	ICE112	1	0	0	0
POP114	Nie1-2	ICE127	1	0	0	0
POP115	Nie1-2	ICE134	1	0	0	0
POP116	Nie1-2	ICE173	1	0	0	0
POP117	Nie1-2	ICE216	1	0	0	0
POP118	Nie1-2	Fei-0	1	0	0	0
POP119	Nie1-2	Kastel-1	1	0	0	0
POP120	Nie1-2	Leo-1	1	0	0	0
POP121	Nie1-2	Lerik1-3	1	0	0	0
POP122	Nie1-2	Rus3-1-31	1	0	0	0
POP123	Nie1-2	TueSB30-3	1	0	0	0
POP124	Nie1-2	TueWa1-2	1	0	0	0

POP125	Nie1-2	Xan-1	1	0	0	0
POP128	ICE1	ICE60	1	0	1	0
POP129	ICE1	ICE138	1	1	1	0
POP130	ICE1	ICE152	1	0	1	0
POP131	ICE1	ICE163	1	0	1	0
POP132	ICE1	Nie1-2	0	0	0	0
POP133	ICE1	Yeg-1	1	0	0	0
POP134	ICE7	ICE102	1	0	0	0
POP135	ICE7	ICE106	1	0	1	0
POP136	ICE7	Don-0	1	0	1	0
POP137	ICE7	Fei-0	0	0	0	0
POP138	ICE7	TueWa1-2	1	0	1	0
POP139	ICE21	ICE213	1	0	0	0
POP140	ICE21	ICE216	1	0	0	0
POP141	ICE21	Ey15-2	1	0	0	0
POP142	ICE21	Koch-1	1	0	0	0
POP143	ICE21	Star-8	1	0	0	0
POP144	ICE21	WalhaesB4	1	0	0	0
POP145	ICE21	Yeg-1	1	0	0	0
POP146	ICE29	ICE63	1	0	1	0
POP147	ICE29	ICE92	1	0	0	0
POP148	ICE29	ICE150	0	0	0	0
POP149	ICE29	Dal-10	1	0	0	0
POP150	ICE29	Qui-0	1	0	0	0
POP151	ICE29	Star-8	1	1	1	0
POP152	ICE29	WalhaesB4	0	0	0	0
POP153	ICE50	ICE1	1	0	0	0
POP154	ICE50	ICE79	0	0	0	0
POP155	ICE50	ICE153	0	0	0	0
POP155.2	ICE50	ICE153	1	0	1	0
POP156	ICE50	ICE163	1	0	0	0
POP157	ICE50	Bak-2	0	0	0	0
POP158	ICE50	Dog-4	0	0	0	0
POP159	ICE50	Ey15-2	1	0	0	0
POP160	ICE50	Mer-6	1	0	1	0
POP161	ICE60	ICE1	1	1	1	0
POP162	ICE60	ICE49	0	0	0	0
POP163	ICE60	ICE98	0	0	0	0
POP164	ICE60	ICE104	1	0	0	0
POP165	ICE60	ICE138	0	0	0	0
POP166	ICE60	Agu-1	0	0	0	0
POP167	ICE60	Lerik1-3	0	0	0	0

POP168	ICE60	Yeg-1	1	0	0	0	0
POP169	ICE61	ICE106	1	0	0	0	0
POP170	ICE61	ICE127	1	0	0	0	0
POP171	ICE61	ICE150	1	0	0	0	0
POP172	ICE61	Fei-0	1	0	0	0	0
POP173	ICE61	Lenk1-3	1	0	0	0	0
POP174	ICE61	Pra-6	0	0	0	0	0
POP175	ICE61	Qui-0	1	1	1	1	0
POP176	ICE70	ICE49	1	0	0	0	0
POP177	ICE70	ICE50	1	0	0	0	0
POP178	ICE70	ICE98	1	0	1	0	0
POP179	ICE70	Cdm-0	1	0	1	0	0
POP180	ICE70	TueSB30-3	1	0	0	0	0
POP181	ICE72	ICE71	0	0	0	0	0
POP182	ICE72	ICE75	0	0	0	0	0
POP183	ICE72	ICE111	1	0	1	0	0
POP184	ICE72	ICE228	0	0	0	0	0
POP185	ICE72	Agu-1	1	0	0	0	0
POP186	ICE72	Leo-1	1	0	0	0	0
POP187	ICE72	WalhaesB4	1	0	0	0	0
POP188	ICE72	Xan-1	0	0	0	0	0
POP189	ICE73	ICE49	1	0	0	0	0
POP190	ICE73	ICE97	0	0	0	0	0
POP191	ICE73	ICE106	1	0	0	0	0
POP192	ICE73	ICE212	0	0	0	0	0
POP193	ICE73	Bak-2	0	0	0	0	0
POP194	ICE73	Fai-0	1	0	0	0	0
POP195	ICE73	Nia1-2	1	0	0	0	0
POP196	ICE73	Ped-0	0	0	0	0	0
POP197	ICE75	ICE60	0	0	0	0	0
POP198	ICE75	ICE91	1	1	1	1	0
POP199	ICE75	ICE119	1	0	0	0	0
POP200	ICE75	ICE150	0	0	0	0	0
POP201	ICE75	Bak-7	1	1	1	1	0
POP202	ICE75	Cdm-0	1	0	0	0	0
POP203	ICE79	ICE93	1	0	0	0	0
POP204	ICE79	ICE98	1	1	1	1	0
POP205	ICE79	ICE104	0	0	0	0	0
POP206	ICE79	ICE153	1	1	1	1	0
POP207	ICE79	Ey15-2	1	0	0	0	0
POP208	ICE79	Koch-1	1	0	0	0	0
POP209	ICE79	Nie1-2	1	0	0	0	0

POP210	ICE79	TueV13	0	0	0	0	0
POP211	ICE79	WalhaesB4	1	0	0	0	0
POP212	ICE91	ICE7	1	1	1	1	0
POP213	ICE91	ICE33	0	0	0	0	0
POP214	ICE91	Ey15-2	1	0	0	0	0
POP215	ICE91	Fei-0	1	0	0	0	0
POP216	ICE91	HKT2.4	1	0	0	0	0
POP217	ICE91	Nie1-2	0	0	0	0	0
POP218	ICE91	TueSB30-3	1	0	1	0	0
POP219	ICE91	TueV13	0	0	0	0	0
POP220	ICE92	ICE49	1	0	0	0	0
POP221	ICE92	ICE60	1	0	0	0	0
POP222	ICE92	ICE75	1	0	0	0	0
POP223	ICE92	ICE130	1	0	0	0	0
POP224	ICE92	ICE134	1	0	0	0	0
POP225	ICE92	Ey15-2	1	0	0	0	0
POP226	ICE93	ICE7	1	0	0	0	0
POP227	ICE93	ICE102	1	0	0	0	0
POP228	ICE93	ICE106	1	0	0	0	0
POP229	ICE93	ICE112	1	0	0	0	0
POP230	ICE93	ICE127	1	0	0	0	0
POP231	ICE93	ICE173	1	0	0	0	0
POP232	ICE93	ICE216	0	0	0	0	0
POP233	ICE93	Star-8	1	0	0	0	0
POP234	ICE97	ICE29	1	0	0	0	0
POP235	ICE97	ICE36	1	0	1	0	0
POP236	ICE97	ICE130	1	0	0	0	0
POP237	ICE97	ICE181	1	0	1	0	0
POP238	ICE97	ICE216	1	0	0	0	0
POP239	ICE97	Nemrut-1	1	0	0	0	0
POP240	ICE97	Qui-0	1	0	0	0	0
POP241	ICE97	Sha	1	0	0	0	0
POP242	ICE98	ICE29	1	0	0	0	0
POP243	ICE98	ICE104	1	0	0	0	0
POP244	ICE98	ICE107	1	0	0	0	0
POP245	ICE98	ICE228	1	0	0	0	0
POP246	ICE98	Bak-7	1	0	0	0	0
POP247	ICE98	Mer-6	1	0	0	0	0
POP248	ICE98	Vash-1	1	0	0	0	0
POP249	ICE98	Yag-1	1	0	0	0	0
POP250	ICE102	ICE21	1	0	0	0	0
POP251	ICE102	ICE29	1	0	0	0	0

POP252	ICE102	ICE98	1	0	0	0	0
POP253	ICE102	ICE111	1	0	0	0	0
POP254	ICE102	ICE138	0	0	0	0	0
POP255	ICE102	ICE212	1	0	0	0	0
POP256	ICE102	Nia1-2	1	0	1	0	0
POP257	ICE104	ICE21	1	0	0	0	0
POP258	ICE104	ICE70	1	0	0	0	0
POP259	ICE104	ICE71	0	0	0	0	0
POP260	ICE104	ICE153	0	0	0	0	0
POP261	ICE104	ICE226	1	0	0	0	0
POP262	ICE104	Sha	1	0	0	0	0
POP263	ICE104	Star-8	0	0	0	0	0
POP264	ICE104	TueWat1-2	1	0	0	0	0
POP265	ICE106	ICE36	0	0	0	0	0
POP266	ICE106	ICE50	0	0	0	0	0
POP267	ICE106	ICE61	0	0	0	0	0
POP268	ICE106	ICE93	0	0	0	0	0
POP269	ICE106	ICE130	0	0	0	0	0
POP270	ICE106	ICE138	0	0	0	0	0
POP271	ICE106	ICE153	1	0	0	0	0
POP272	ICE106	ICE226	1	0	0	0	0
POP273	ICE107	ICE130	1	0	0	0	0
POP274	ICE107	ICE173	1	0	0	0	0
POP275	ICE107	ICE216	0	0	0	0	0
POP276	ICE107	Bak-7	1	0	1	0	0
POP277	ICE107	Lag2.2	0	0	0	0	0
POP278	ICE107	Leo-1	1	0	0	0	0
POP279	ICE107	Nemrut-1	1	0	0	0	0
POP280	ICE107	TueV13	0	0	0	0	0
POP281	ICE111	ICE7	0	0	0	0	0
POP282	ICE111	ICE21	1	0	0	0	0
POP283	ICE111	ICE63	1	0	0	0	0
POP284	ICE111	ICE104	0	0	0	0	0
POP285	ICE111	ICE112	0	0	0	0	0
POP286	ICE111	ICE134	0	0	0	0	0
POP287	ICE111	WalhaesB4	1	0	0	0	0
POP287.2	ICE111	WalhaesB4	0	0	0	0	0
POP288	ICE119	ICE21	1	0	0	0	0
POP289	ICE119	ICE50	1	0	0	0	0
POP290	ICE119	ICE60	1	0	0	0	0
POP291	ICE119	ICE134	1	0	1	0	0
POP292	ICE119	Bak-7	1	0	0	0	0

POP293	ICE119	Cdm-0	0	0	0	0	0
POP294	ICE119	Del-10	1	0	0	0	0
POP295	ICE119	Ey15-2	1	0	0	0	0
POP296	ICE119	HKT2.4	0	0	0	0	0
POP297	ICE120	ICE60	0	0	0	0	0
POP298	ICE120	ICE70	0	0	0	0	0
POP299	ICE120	ICE71	0	0	0	0	0
POP300	ICE120	ICE73	0	0	0	0	0
POP301	ICE120	ICE106	0	0	0	0	0
POP302	ICE120	ICE112	1	0	1	0	0
POP303	ICE120	Istau-1	0	0	0	0	0
POP304	ICE120	Lag2.2	1	0	0	0	0
POP305	ICE130	ICE102	0	0	0	0	0
POP306	ICE130	ICE119	1	0	0	0	0
POP307	ICE130	Lag2.2	1	0	1	0	0
POP308	ICE130	Leo-1	1	0	0	0	0
POP309	ICE130	Ped-0	0	0	0	0	0
POP310	ICE130	Qui-0	1	0	0	0	0
POP311	ICE130	Vash-1	1	0	0	0	0
POP312	ICE130	WalhaesB4	1	0	1	0	0
POP313	ICE134	ICE21	1	0	1	0	0
POP314	ICE134	ICE71	1	0	1	0	0
POP315	ICE134	ICE73	1	0	0	0	0
POP316	ICE134	ICE152	1	0	1	0	0
POP317	ICE134	HKT2.4	1	0	0	0	0
POP318	ICE134	Mer-6	0	0	0	0	0
POP319	ICE134	Xan-1	1	0	0	0	0
POP320	ICE134	Yag-1	1	0	0	0	0
POP321	ICE138	ICE36	1	0	1	0	0
POP322	ICE138	ICE71	1	0	0	0	0
POP323	ICE138	ICE93	1	0	0	0	0
POP324	ICE138	ICE104	1	0	0	0	0
POP325	ICE138	ICE226	1	0	0	0	0
POP326	ICE138	Nemrut-1	1	0	0	0	0
POP327	ICE138	Ped-0	1	0	0	0	0
POP328	ICE138	Star-8	1	0	1	0	0
POP329	ICE150	ICE91	1	0	0	0	0
POP330	ICE150	ICE97	1	0	0	0	0
POP331	ICE150	Agu-1	1	0	1	0	0
POP332	ICE150	HKT2.4	1	0	0	0	0
POP333	ICE150	Lag2.2	1	0	0	0	0
POP334	ICE150	Leo-1	1	0	0	0	0

POP335	ICE150	Vash-1	1	0	0	0	0
POP336	ICE150	Xan-1	1	0	0	0	0
POP337	ICE152	ICE97	1	0	0	0	0
POP338	ICE152	ICE163	1	0	0	0	0
POP339	ICE152	ICE173	1	0	0	0	0
POP340	ICE152	ICE228	1	0	0	0	0
POP341	ICE152	Rue3-1-31	1	0	0	0	0
POP342	ICE152	Star-8	1	1	1	0	0
POP343	ICE152	TueWa1-2	1	0	0	0	0
POP344	ICE152	Yeg-1	1	0	0	0	0
POP345	ICE153	ICE1	1	1	1	0	0
POP346	ICE153	ICE73	1	0	1	0	0
POP347	ICE153	ICE79	1	0	0	0	0
POP348	ICE153	ICE111	1	0	0	0	0
POP349	ICE153	ICE127	1	0	0	0	0
POP350	ICE153	Mer-6	0	0	0	0	0
POP351	ICE153	TueV13	1	1	1	0	0
POP352	ICE153	Vash-1	1	0	1	0	0
POP353	ICE173	ICE7	1	0	0	0	0
POP354	ICE173	ICE50	1	0	1	0	0
POP355	ICE173	ICE70	1	0	0	0	0
POP356	ICE173	ICE104	1	0	0	0	0
POP357	ICE173	ICE111	1	0	0	0	0
POP358	ICE173	ICE150	1	0	0	0	0
POP359	ICE173	Lag2.2	1	0	0	0	0
POP360	ICE173	Star-8	1	0	1	0	0
POP361	ICE181	ICE60	1	0	1	0	0
POP362	ICE181	ICE61	1	0	0	0	0
POP363	ICE181	ICE71	1	0	1	0	0
POP364	ICE181	ICE120	1	0	0	0	0
POP365	ICE181	ICE152	1	1	1	0	0
POP366	ICE181	ICE153	1	1	1	0	0
POP367	ICE181	ICE212	1	0	1	0	0
POP368	ICE181	Nie1-2	1	0	0	0	0
POP369	ICE212	ICE104	0	0	0	0	0
POP370	ICE212	ICE106	1	0	0	0	0
POP371	ICE212	ICE119	1	0	0	0	0
POP372	ICE212	ICE127	1	0	0	0	0
POP373	ICE212	ICE150	1	0	0	0	0
POP374	ICE212	Ey15-2	1	0	0	0	0
POP375	ICE212	HKT2.4	1	0	1	0	0
POP376	ICE212	Lenk1-3	1	0	0	0	0

POP377	ICE213	ICE21	1	0	0	0	0
POP378	ICE213	ICE29	1	0	0	0	0
POP379	ICE213	ICE79	1	0	0	0	0
POP380	ICE213	ICE106	1	0	0	0	0
POP381	ICE213	ICE111	1	0	0	0	0
POP382	ICE213	ICE130	0	0	0	0	0
POP383	ICE213	Ey15-2	1	0	0	0	0
POP384	ICE216	ICE36	1	0	0	0	0
POP385	ICE216	ICE79	1	0	0	0	0
POP386	ICE216	ICE102	1	0	0	0	0
POP387	ICE216	ICE104	1	0	0	0	0
POP388	ICE216	ICE107	1	0	0	0	0
POP389	ICE216	ICE153	1	0	0	0	0
POP390	ICE216	Kastel-1	1	0	0	0	0
POP391	ICE216	Ped-0	1	0	0	0	0
POP392	ICE228	ICE73	0	0	0	0	0
POP393	ICE228	ICE152	1	0	0	0	0
POP394	ICE228	ICE163	1	1	1	0	0
POP395	ICE228	Bak-2	1	0	0	0	0
POP396	ICE228	HKT2.4	1	0	0	0	0
POP397	ICE228	Lag2.2	1	1	1	0	0
POP398	ICE228	Nie1-2	1	0	0	0	0
POP399	Agu-1	ICE29	1	0	0	0	0
POP400	Agu-1	ICE150	1	0	1	0	0
POP401	Agu-1	ICE153	1	0	1	0	0
POP402	Agu-1	ICE163	1	0	0	0	0
POP403	Agu-1	Bak-2	1	0	0	0	0
POP404	Agu-1	Ey15-2	1	0	0	0	0
POP405	Agu-1	Yeg-1	0	0	0	0	0
POP406	Bak-7	ICE71	0	0	0	0	0
POP407	Bak-7	ICE152	1	0	0	0	0
POP408	Bak-7	ICE163	0	0	0	0	0
POP409	Bak-7	Bak-2	0	0	0	0	0
POP410	Bak-7	Kastel-1	1	0	0	0	0
POP411	Bak-7	Leo-1	1	0	0	0	0
POP412	Bak-7	Ped-0	0	0	0	0	0
POP413	Bak-7	Vash-1	1	0	0	0	0
POP414	Cdm-0	ICE93	1	0	0	0	0
POP415	Cdm-0	ICE104	1	0	0	0	0
POP416	Cdm-0	ICE120	1	0	0	0	0
POP417	Cdm-0	ICE163	1	0	0	0	0
POP418	Cdm-0	ICE169	1	0	0	0	0

POP419	Cdm-0	Mer-6	1	0	0	0	0
POP420	Cdm-0	Nie1-2	1	0	0	0	0
POP421	Cdm-0	TueWa1-2	1	0	0	0	0
POP422	Del-10	ICE36	0	0	0	0	0
POP423	Del-10	ICE75	1	0	0	0	0
POP424	Del-10	ICE79	1	0	0	0	0
POP425	Del-10	ICE97	1	0	0	0	0
POP426	Del-10	Don-0	1	0	0	0	0
POP427	Del-10	Fei-0	1	0	0	0	0
POP428	Del-10	Ped-0	1	0	0	0	0
POP429	Del-10	Yeg-1	1	0	1	0	0
POP430	Dog-4	ICE29	1	0	0	0	0
POP431	Dog-4	ICE33	1	0	0	0	0
POP432	Dog-4	ICE97	1	0	0	0	0
POP433	Dog-4	ICE119	1	0	0	0	0
POP434	Dog-4	ICE138	1	0	0	0	0
POP435	Dog-4	Agu-1	1	0	0	0	0
POP436	Dog-4	Ey15-2	1	0	0	0	0
POP437	Dog-4	Sha	1	0	0	0	0
POP438	Don-0	ICE70	1	0	0	0	0
POP439	Don-0	ICE71	1	0	0	0	0
POP440	Don-0	ICE75	1	0	1	0	0
POP441	Don-0	ICE79	1	0	0	0	0
POP442	Don-0	ICE102	1	1	1	0	0
POP443	Don-0	ICE107	1	0	1	0	0
POP444	Don-0	ICE138	1	0	0	0	0
POP445	Don-0	ICE163	1	0	0	0	0
POP446	Fei-0	ICE21	0	0	0	0	0
POP447	Fei-0	ICE49	1	0	0	0	0
POP448	Fei-0	ICE50	1	0	0	0	0
POP449	Fei-0	ICE169	1	0	0	0	0
POP450	Fei-0	ICE228	1	0	0	0	0
POP451	Fei-0	Bak-7	0	0	0	0	0
POP452	Fei-0	Cdm-0	1	0	0	0	0
POP453	Fei-0	Ey15-2	1	0	0	0	0
POP454	HKT2.4	ICE49	1	0	1	0	0
POP455	HKT2.4	ICE60	1	0	0	0	0
POP456	HKT2.4	ICE63	1	0	0	0	0
POP457	HKT2.4	Cdm-0	1	0	0	0	0
POP458	HKT2.4	Del-10	1	0	0	0	0
POP459	HKT2.4	Istisu-1	1	0	0	0	0
POP460	HKT2.4	Rue3-1-31	1	0	0	0	0

POP461	HKT2.4	Vie-0	1	0	0	0	0
POP462	Istisu-1	ICE29	1	1	1	0	0
POP463	Istisu-1	ICE92	1	0	0	0	0
POP464	Istisu-1	ICE169	1	0	0	0	0
POP465	Istisu-1	ICE212	1	1	1	0	0
POP466	Istisu-1	ICE213	1	0	0	0	0
POP467	Istisu-1	Lag2.2	1	0	0	0	0
POP468	Istisu-1	Nie1-2	0	0	0	0	0
POP469	Koch-1	ICE7	1	0	0	0	0
POP470	Koch-1	ICE33	1	0	0	0	0
POP471	Koch-1	ICE98	1	0	0	0	0
POP472	Koch-1	ICE119	1	0	0	0	0
POP473	Koch-1	ICE134	1	0	0	0	0
POP474	Koch-1	ICE213	1	0	0	0	0
POP475	Koch-1	Koch-1	0	0	0	0	0
POP476	Koch-1	WalhaesB4	1	0	0	0	0
POP476.2	Koch-1	WalhaesB4	0	0	0	0	0
POP477	Mer-6	ICE49	1	0	0	0	0
POP478	Mer-6	ICE60	0	0	0	0	0
POP479	Mer-6	ICE91	1	0	0	0	0
POP480	Mer-6	ICE98	1	0	0	0	0
POP480.2	Mer-6	ICE98	0	0	0	0	0
POP481	Mer-6	Cdm-0	1	0	0	0	0
POP482	Mer-6	Fei-0	1	0	0	0	0
POP483	Mer-6	TueWa1-2	1	1	1	0	0
POP484	Mer-6	WalhaesB4	1	0	0	0	0
POP485	Nemrut-1	ICE21	1	0	0	0	0
POP486	Nemrut-1	ICE79	1	1	1	0	0
POP487	Nemrut-1	ICE92	1	0	0	0	0
POP488	Nemrut-1	ICE212	0	0	0	0	0
POP488.2	Nemrut-1	ICE212	1	0	1	0	0
POP489	Nemrut-1	Leo-1	1	0	0	0	0
POP490	Nemrut-1	Nemrut-1	0	0	0	0	0
POP491	Nemrut-1	Nie1-2	1	0	0	0	0
POP492	Pra-6	ICE79	1	0	1	0	0
POP493	Pra-6	Agu-1	1	1	1	0	0
POP494	Pra-6	Fei-0	1	0	0	0	0
POP495	Pra-6	Kastel-1	1	0	0	0	0
POP496	Pra-6	Tuescha9	1	0	0	0	0
POP497	Pra-6	Vie-0	0	0	0	0	0
POP498	Pra-6	Xan-1	1	0	1	0	0
POP499	Pra-6	Yeg-1	1	0	0	0	0

POP500	Qui-0	ICE106	0	0	0	0	0
POP501	Qui-0	ICE181	0	0	0	0	0
POP502	Qui-0	Don-0	1	0	0	1	0
POP503	Qui-0	Rue3-1-31	1	0	0	0	0
POP504	Qui-0	Star-8	1	0	0	0	0
POP505	Qui-0	Tuescha9	1	0	0	1	0
POP506	Qui-0	TueWa1-2	1	0	0	0	0
POP507	Rue3-1-31	ICE29	1	0	0	0	0
POP508	Rue3-1-31	ICE60	0	0	0	0	0
POP509	Rue3-1-31	ICE97	1	0	0	0	0
POP510	Rue3-1-31	ICE153	1	0	0	0	0
POP511	Rue3-1-31	ICE163	1	0	0	0	0
POP512	Rue3-1-31	Dag-4	1	0	0	0	0
POP513	Rue3-1-31	Lag2.2	1	0	0	0	0
POP514	Rue3-1-31	Lerk1-3	1	0	0	0	0
POP515	Sha	ICE49	1	1	1	1	0
POP516	Sha	ICE106	1	0	0	0	0
POP517	Sha	Ey15-2	1	1	1	1	0
POP518	Sha	Lerk1-3	1	0	0	0	0
POP519	Sha	Nie1-2	1	0	0	0	0
POP520	Sha	Rue3-1-31	1	0	0	0	0
POP521	Sha	Star-8	1	1	1	1	0
POP522	Sha	TueWa1-2	1	0	0	0	0
POP523	Star-8	ICE49	1	0	0	0	0
POP524	Star-8	ICE104	1	1	1	1	0
POP525	Star-8	ICE106	1	1	1	1	0
POP526	Star-8	ICE130	1	1	1	1	0
POP527	Star-8	ICE228	1	0	1	0	0
POP528	Star-8	Cdm-0	1	0	0	0	0
POP529	Star-8	Fei-0	1	0	1	0	0
POP530	Star-8	Vie-0	1	1	1	1	0
POP531	TueSB30-3	ICE49	1	0	0	1	0
POP532	TueSB30-3	ICE61	1	0	0	0	0
POP533	TueSB30-3	ICE138	1	0	0	0	0
POP534	TueSB30-3	ICE152	1	0	0	0	0
POP535	TueSB30-3	Istiau-1	1	0	0	0	0
POP536	TueSB30-3	Nemrut-1	1	0	0	0	0
POP537	TueSB30-3	Pra-6	0	0	0	0	0
POP538	TueSB30-3	Qui-0	1	0	0	0	0
POP539	Tuescha9	ICE63	1	0	0	0	0
POP540	Tuescha9	ICE97	1	0	0	0	0
POP541	Tuescha9	ICE119	1	0	0	0	0

POP542	Tuescha9	ICE138	0	0	0	0	0
POP543	Tuescha9	ICE181	1	0	0	1	0
POP544	Tuescha9	ICE213	1	0	0	0	0
POP545	Tuescha9	Bak-2	1	0	0	0	0
POP546	Tuescha9	Koch-1	1	0	0	0	0
POP547	TueV13	ICE63	1	1	1	1	0
POP548	TueV13	ICE93	1	0	0	0	0
POP549	TueV13	ICE120	1	0	0	0	0
POP550	TueV13	ICE153	1	0	0	0	0
POP551	TueV13	Ey15-2	1	0	0	0	0
POP552	TueV13	Istiau-1	1	0	0	0	0
POP553	TueV13	Lerk1-3	1	0	0	0	0
POP554	TueV13	Pa-d-0	1	0	0	0	0
POP555	TueWa1-2	ICE60	1	0	0	0	0
POP556	TueWa1-2	ICE70	1	0	0	0	0
POP557	TueWa1-2	ICE71	1	0	0	0	0
POP558	TueWa1-2	ICE119	1	0	0	0	0
POP559	TueWa1-2	ICE130	1	0	0	0	0
POP560	TueWa1-2	ICE228	1	0	0	0	0
POP561	TueWa1-2	Fei-0	0	0	0	0	0
POP561.2	TueWa1-2	Fei-0	0	0	0	0	0
POP562	TueWa1-2	Vie-0	1	0	0	0	0
POP563	Vie-0	ICE93	1	0	0	0	0
POP564	Vie-0	ICE106	1	0	0	0	0
POP565	Vie-0	ICE138	1	0	0	0	0
POP566	Vie-0	ICE169	1	0	0	0	0
POP567	Vie-0	ICE216	1	0	0	0	0
POP568	Vie-0	Bak-2	1	0	0	0	0
POP569	Vie-0	Bak-7	1	0	0	0	0
POP570	Vie-0	Yeg-1	0	0	0	0	0
POP571	WalhaesB4	ICE61	1	0	0	0	0
POP572	WalhaesB4	ICE75	1	0	0	0	0
POP573	WalhaesB4	ICE152	1	0	0	0	0
POP574	WalhaesB4	Cdm-0	1	0	0	1	0
POP575	WalhaesB4	Nie1-2	1	0	0	0	0
POP576	WalhaesB4	Star-8	1	0	0	0	0
POP577	WalhaesB4	TueSB30-3	1	0	0	0	0
POP578	Yeg-1	ICE71	1	0	0	0	0
POP579	Yeg-1	ICE97	1	0	0	0	0
POP580	Yeg-1	ICE106	1	0	0	0	0
POP581	Yeg-1	Lao-1	1	0	0	0	0
POP582	Yeg-1	Lerk1-3	1	0	0	0	0

POP583	Yeg-1	Nie1-2	1	0	0	0	0
POP584	Yeg-1	Vie-0	0	0	0	0	0
POP585	Yeg-1	WalhaesB4	1	0	0	0	0

Table 2. Candidate intervals for distorted loci.

<i>Interval refinement method</i>	<i>Population ID or Shared grandparent ID</i>	<i>Distorted chromosome</i>	<i>Interval start (Mb)</i>	<i>Interval stop (Mb)</i>	<i>Interval length (Mb)</i>
K-mer analysis	POP007	Chromosome 1	20.000	27.750	7.750
K-mer analysis	POP026	Chromosome 5	11.300	15.050	3.750
K-mer analysis	POP035	Chromosome 1	13.350	17.950	4.600
K-mer analysis	POP063	Chromosome 3	1.700	7.350	5.650
K-mer analysis	POP064	Chromosome 5	NA	NA	NA
K-mer analysis	POP100	Chromosome 2	3.550	5.000	1.450
Bulk segregant analysis	Star-8	Chromosome 1	8.553	10.591	2.038
Bulk segregant analysis	ICE49	Chromosome 1	21.414	26.362	4.948
Bulk segregant analysis	ICE63	Chromosome 1	8.254	25.007	16.753

7. Discussion

As a complement to the topic focused discussions contained in the respective manuscripts, I would like to discuss how the results of each research project are linked to the evolution of genome architecture while exploring remaining questions and future research directions.

1. The evolution of DNA methylation in the Brassicaceae

This comparative genomics experiment was designed to address two lines of inquiry. First, I sought to characterize the extent of between-species variation in DNA methylation. The second line of inquiry focused on exploring the magnitude of within-species DNA methylation variation in response to various stimuli. The major advance of this work came from the ability to directly compare, or align, a large proportion of the focal species' genomes. These alignments were facilitated by the relatively recent divergence between *Capsella rubella*, *Arabidopsis lyrata*, and *Arabidopsis thaliana*, which occurred less than 20 million years ago [86]. Three major conclusions resulted from this research and these conclusions revealed two evolutionary modes that facilitate the evolution of DNA methylation in plant genomes.

The majority of DNA methylation is driven by interspecific TE evolution

The first key conclusion of this work is that the majority of DNA methylation is linked to transposable elements. As a result, the largest source of interspecific variation in DNA methylation is the rapid evolution of these elements. In all species, most methylated nucleotides are found in local groups, or regions, which were computationally identified using a hidden Markov model. Together these regions account for 15 to 30% of the studied genomes and are heavily enriched for TE sequences. The general chromosomal distributions of TEs and associated DNA methylation are comparable across species; in all cases these features are enriched in centromeric and pericentromeric genomic regions. Despite this similarity, over 95% of methylated regions are not contained within the multi-species genomic alignments, which cover roughly two-thirds of the *A. thaliana* genome. This

finding suggests that DNA methylation is predominantly a phenotype and that the evolution of this mark is driven by the forces that shape the underlying genetic sequence.

The loss of three ancestral centromeres on the path leading to the *A. thaliana* lineage provides support for the rapid turnover of repeat sequences and associated epigenetic marks. Chromosome fusion, inversions, and translocations have reorganized the ancestral Brassicaceae karyotype to give rise to the five derived chromosomes found in *A. thaliana* [46, 87]. This large-scale genomic rearrangement has placed ancestral centromeres in regions of elevated recombination in *A. thaliana*. Presumably as a result of elevated recombination and the increased efficacy of selection in these regions, TE sequences have been purged from ancestral centromeres resulting in euchromatic levels of TE and DNA methylation densities. The genomic response to massive shifts in genomic architecture only took place in the time since *A. thaliana* diverged from its congener, *A. lyrata*, 10 million years ago [86].

The decreased density of TEs in gene-rich, euchromatic sequences in most species indicates either the TE insertions are biased, preferentially inserting into heterochromatic regions, or that new insertions are purged from these functional regions. That *A. lyrata* has experienced a recent burst of transposition, occurring only 0.6 million years ago [46], and that this species is enriched for euchromatic, methylated TEs provides evidence for secondary loss of these sequences. In many plant species, methylated TEs located in gene neighborhoods can dampen transcription of endogenous genes [51, 52] and TE-associated silencing of linked loci has resulted in phenotypic perturbations [88-91]. The role of methylation, and TEs in general, in shaping phenotypic diversity has long interested researchers. Although there is evidence that such loci may generate adaptive phenotypic variation, such as those that have been selected during maize domestication [92], the majority of methylated TEs in euchromatic regions are likely deleterious [48]. This has been demonstrated in *A. thaliana*, where TE insertions occur at low frequencies in surveyed populations and inserted regions display evidence of purifying selection [48]. Together, this indicates that while methylated TEs can perturb transcription, these changes are frequently deleterious, and when

such changes reside in highly recombining, euchromatic sequences they can be swiftly purged from the genome. Given that gene expression profiles are conserved across millions of years of evolution [93], any phenotypic consequences derived from TE-driven shifts in DNA methylation are predominantly transitory.

Exon-linked DNA methylation is conserved over long periods

In contrast to TE methylation, exon-linked DNA methylation, known as gene body methylation, is conserved over long evolutionary periods, indicative of differing evolutionary forces acting on gene body methylation and TE or intergenic methylation [94, 95]. Methylation in exons is mostly restricted to symmetrical CG sites and has been shown to correlate with transcription levels [35, 36, 96, 97]. Although roughly two-thirds of protein coding genes are exon-methylated, these sites only account for a fraction of genome-wide DNA methylation. Prior to this work, the conservation of DNA methylation had been demonstrated in grass species diverged by 53 million years. Intraspecific correlation of gene body methylation levels over such long periods is indicative of selective constraint [94]. Using genomic alignments, I showed that not only are average gene body methylation levels conserved, but also that a large fraction of exon-linked CG sites are methylated in all three species, indicating that selective constraint preserves DNA methylation at specific nucleotides. However, it is still unclear if selection is directly maintaining CG methylation or if correlated methylation rates are the result of selection acting on another biological process, for example transcription or chromatin compaction, which in turn affects DNA methylation. One possibility is that exon methylation is the result of intron-exon structure. DNA methylation has been shown to correlate with nucleosome positioning and these protein complexes are carefully positioned to delineate exon-intron boundaries [96, 98]. In the focal species, CG methylation is enriched in the center of exons and most heavily conserved when exon lengths are constant across species. Further work is needed to dissect the causal process driving the preservation of gene body methylation levels, but I hypothesize that purifying selection is acting on gene expression levels, which in turn maintains a chromatin environment required for transcription. Instead of purifying selection acting to

purge methylated sequences, this mode of selection is acting to maintain CG methylation levels in exons for millions of years, possibly as a byproduct of transcription.

Tissue-specific TE regulation gives rise to intraspecific methylation variation

The final conclusion of this work is that DNA methylation responds to both developmental and environmental cues. In all three species, nearly 10% of methylated cytosines varied significantly in root-shoot comparisons while about 1% varied between temperature treatments. Despite the 10-fold difference in DNA methylation response, the magnitude of transcriptional response was comparable, with differential regulation occurring in over 4,000 genes in both cases. Importantly, we identified no global correlation between differential DNA methylation and differential gene expression. Instead, we found that DNA methylation in transposons and also in gene bodies is elevated in shoot tissues in general. This suggests that deregulation of global DNA methylation levels in the root may be the driving source of tissue-specific variation in DNA methylation. Plants, unlike animals, do not partition their germ cells early in development. Instead, the plant germline is designated from meristematic tissue much later in development. As a result, tight control of TE sequences is necessary in the shoot meristems to minimize TE proliferation and associated mutagenic effects [99]. In summary, the host's requirement to minimize insertion events in germ cells has given rise to global shifts in DNA methylation which do not seem to coordinate transcriptional changes between the root and shoot.

Conclusion

Overall, the majority of DNA methylation is linked to TEs, an important driver of variation in both genome size and genome architecture. As a result of this linkage, DNA methylation is enriched in non-recombining, centromeric regions and purged from euchromatic, gene-dense sequences. The fast evolution of such sequences indicates that any transcriptional consequences are short-lived, unless they are replaced by genetically encoded regulation. Despite empirical examples of methylation induced phenotypic shifts, the contribution of this epigenetic mark to lasting phenotypic variation is restricted

by selection acting to purge underlying transposable elements. While this is true for the vast majority of TEs, beneficial TE insertions will be retained for longer periods of time.

II. The genetic architecture of non-additive hybrid phenotypes

The aim of this large-scale phenotyping experiment in *Arabidopsis thaliana* was to address which types of genetic variants give rise to phenotypic variation, a question that has interested evolutionary geneticists for centuries. To dissect the relative contribution of genetic drift, or mutation-selection balance, to phenotypic variation, we chose to use a particular experimental system, F₁ hybrids. There is evidence to suggest that many, deleterious mutations, i.e. those under mutation-selection balance, underlie inbreeding depression (reviewed in [80]), a term that refers to the suite of inferior phenotypes that arise in the inbred progeny of a naturally outbreeding individual. As a result of this work, the biological complement to inbreeding depression, known as heterosis, is thought to be due to the genetic complementation of the same deleterious mutations en route to elimination from the population (reviewed in [80]). Heterosis is rampant in first generation hybrids, making this genetic material particularly relevant for characterizing the influence of genetic drift in driving phenotypic variation. The major advance of this experiment came from the coupling of valuable genetic material, a large collection of F₁ hybrids [77], to genome-wide association mapping. With this material we identified genetic variants associated with hybrid phenotypes suggesting that forces other than genetic drift may be shaping phenotypic variation in *A. thaliana*.

Predicted source of genetic variants associated with hybrid vigor

The mutation-selection balance theory posits that the number of deleterious mutations reaches an equilibrium where their rate of accumulation is equivalent to the rate that they are removed from the population by selection [100]. At any particular snap-shot in evolutionary time each genome will contain a set of these variants in the process of being purged, many of them of small effect. If hybrid phenotypes result from such mutations, the

individual effects will be so small that they will be invisible to association mapping approaches. In fact, we used simulations to show that individual variants that contribute less than 5% to phenotypic variation will not be detected in our population. Even in traditional, biparental QTL mapping studies, these small-effect variants will also not be identified unless large population sizes are used. That we could detect genetic loci associated with hybrid phenotypes indicates the entirety of the genetic architecture is not due to the accumulation of small effect mutations under mutation-selection balance.

Instead, we found at least nine independent loci that are each linked to one or more measured hybrid phenotypes. Together the associated loci can explain up to 40% of the variance of a particular phenotype, a major success for association mapping studies. In addition to uncovering large-effect variants, associated loci displayed both classical dominant and overdominant phenotypic effects. Under the mutation-selection balance hypothesis, both small and large-effect recessive, deleterious loci are constantly purged, ensuring that they remain at low frequencies in outcrossing populations (reviewed in [80]). However, working in an inbreeding species such as *A. thaliana* complicates these expectations. The reduction in effective population size that results from a shift in mating system will reduce the efficacy of purifying selection acting against deleterious loci, increasing their population frequency (reviewed in [80]). If these loci are retained after the shift to inbreeding they will be difficult to remove from the population and these loci will give rise to classical dominant phenotypic effects when complemented in the hybrids. Four of the nine associated loci exhibit dominant effects and may be the result of inefficient purifying selection against recessive, deleterious loci in this inbreeding species.

There are multiple evolutionary scenarios that are used to explain the segregation of overdominant genetic loci in populations. First, overdominant loci, or loci where the heterozygote phenotype is more extreme than either homozygous genotypic class, may be the result of pseudo-overdominance. Pseudo-overdominance describes the situation where two recessive deleterious mutations are physically linked in repulsion and thus cannot be separated easily by recombination (reviewed in [80]). Although the length of

linkage blocks is reduced in large dialleles due to the increase in historical recombination events relative to traditional QTL mapping populations that rely only on modern recombination, the possibility of pseudo-overdominance cannot be discarded. In fact, alleles with overdominant phenotypic effects exhibit longer local linkage disequilibrium than their dominant counterparts. Alternatively, overdominant loci could be due to true overdominance, or the effect of only a single variant (reviewed in [80]). Single mutations can be maintained in the heterozygous state due to balancing selection or antagonistic pleiotropy (reviewed in [80]). In such cases of heterozygote advantage, overdominant mutations will be maintained in populations for long periods of time. If the overdominantly acting loci are instances of true overdominance, then it would suggest a role for selection in shaping the phenotypic variation in hybrids, but molecular dissection of associated loci is necessary to discern between these two possibilities.

Conclusions

That both overdominant and dominant loci were identified in this genome-wide association experiment, suggests that both genetic drift and selection may contribute substantially to hybrid phenotypes. Although large-effect loci were detected, these loci still did not explain the entirety the phenotypic variation. This means that in addition to the nine large-effect loci, many additional small-effect variants may be contributing to the measured phenotypes. As I mentioned before, the novelty of this experiment resides in coupling the diallel crossing scheme to genome-wide association mapping. Despite the success of this approach, increasing the size of the diallel as well as the diversity of the contributing genetic would increase the probability of detecting additional loci underlying heterotic phenotypes. Additionally, further molecular dissection of the large-effect loci would provide a complete picture of the evolutionary forces that give rise to heterosis.

III. A species-wide screen for intraspecific genetic barriers

A large screen for biased transmission of alleles in *A. thaliana* was carried out with the goal of establishing the rate of such events at the species

level. Examples of segregation distortion have been documented in myriad organisms, but whether these events are genetic anomalies or common occurrences is unknown. The first step towards appreciating the contribution of segregation distortion to the formation of intraspecific genetic barriers is to fully characterize the likelihood of its occurrence in genetically diverse germplasm. The major advance of this work came from coupling modern high-throughput sequencing with a large set of genetically diverse segregating F₂ populations. By surveying the extent of biased inheritance in a large set of diverse germplasm, I was able to provide one of the first estimates of the extent of this phenomenon at a species-wide level. The observed rate of segregation distortion (12-24%) indicates that there are many intraspecific genetic barriers segregating in this species. Fine-scale molecular dissection of distorted loci will be necessary in order to appreciate the magnitude of this finding and the actual contribution of such loci to hybrid dysfunction and subsequent speciation.

The genetic architecture of segregation distortion in A. thaliana

The genetic architecture of segregation distortion can hint at the underlying biological process. Typically, segregation distortion results from deleterious epistasis, competition or differential fitness between gametes, or non-random segregation during meiosis (reviewed in [101]). While deleterious epistasis can result from incompatible interactions between two or more loci, only two-locus interactions will sufficiently perturb the expected allele frequency in the surveyed F₂ populations. I found that in a majority of cases, segregation distortion was significant for only a single genomic region, although instances of tightly linked loci cannot be ruled out. This indicates that two-locus deleterious epistasis, such as classical Bateson-Dobzhansky-Muller genetic incompatibility, is not a major driver of the segregation distortion observed here. Previous work has shown that in *Drosophila melanogaster* deleterious epistasis occurs frequently, with each of the eight surveyed strains carrying one or two contributing loci [102]. This discrepancy may be the result of mating system differences between the two systems, with more deleterious mutations maintained in large, outbreeding populations like *D. melanogaster* [103].

Without molecular identification of the causal loci it is difficult to speculate on the relative contribution of gamete competition and non-random meiotic segregation, or meiotic drive, to biased genetic transmission. The few molecularly characterized cases of meiotic drive are caused by the expansion of repeat or heterochromatic content, likely manipulating their affinity for the meiotic machinery [78, 79, 104]. Interestingly, a number of the candidate regions that were identified in this screen span centromeres, at least two of which have been narrowed to intervals of less than 5 Mb. If rapid shifts in repetitive sequences found in centromeres repeatedly spawn loci capable of driving their own inheritance, then there are severe consequences for the fast evolution of genome architecture in plants with important implications for the formation of genetic barriers in this species.

Conclusions

I found that biased transmission of alleles is not a rare event in *A. thaliana* and that up to 24% of segregating populations experience atypical inheritance. Additionally, over 50% of the grandparental accessions contribute alleles that are favored in at least one cross, indicating that the source of these alleles is extensive. To fully appreciate the extent that the evolution of genome architecture obstructs the free flow of genetic information in this species, detailed molecular genetic dissection of each causal locus is imperative. Detailed knowledge of the favored loci will also provide insight into the evolutionary forces driving intraspecific genetic barriers.

8. References

1. Gaut BS, Wright SI, Rizzon C, Dvorak J, Anderson LK (2007) Recombination: an underappreciated factor in the evolution of plant genomes. *Nat Rev Genet* 8(1):77-84.
2. Bennett MD, Leitch IJ (2012) Plant DNA C-values database (release 6.0, Dec. 2012). (<http://www.kew.org/cvalues/>).
3. Gregory TR (2005) The C-value enigma in plants and animals: A review of parallels and an appeal for partnership. *Ann Bot* 95(1):133-46.
4. Heslop-Harrison JS, Schwarzacher T (2011) Organisation of the plant genome in chromosomes. *Plant J* 66(1):18-33.
5. Gregory TR (2005) Animal Genome Size Database. (<http://www.genomesize.com/>).
6. Adams KL, Wendel JF (2005) Polyploidy and genome evolution in plants. *Curr Opin Plant Biol* 8(2):135-41.
7. Cui L, Wall PK, Leebens-Mack JH, Lindsay BG, Soltis DE, Doyle JJ, et al. (2006) Widespread genome duplications throughout the history of flowering plants. *Genome Res* 16(6):738-49.
8. Masterson J (1994) Stomatal size in fossil plants: evidence for polyploidy in majority of angiosperms. *Science* 264(5157):421-4.
9. Blanc G, Wolfe KH (2004) Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell* 16(7):1667-78.
10. Tang H, Wang X, Bowers JE, Ming R, Alam M, Paterson AH (2008) Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res* 18(12):1944-54.
11. Van de Peer Y, Maere S, Meyer A (2009) The evolutionary significance of ancient genome duplications. *Nat Rev Genet* 10(10):725-32.
12. Otto SP (2007) The evolutionary consequences of polyploidy. *Cell* 131(3):452-62.
13. Comai L (2005) The advantages and disadvantages of being polyploid. *Nat Rev Genet* 6(11):836-46.
14. Soltis DE, Albert VA, Leebens-Mack J, Bell CD, Paterson AH, Zheng C, et al. (2009) Polyploidy and angiosperm diversification. *Am J Bot* 96(1):336-48.
15. Madlung A, Tyagi AP, Watson B, Jiang H, Kagechi T, Doerge RW, et al. (2005) Genomic changes in synthetic *Arabidopsis* polyploids. *Plant J* 41(2):221-30.
16. Yant L, Hollister JD, Wright KM, Arnold BJ, Higgins JD, Franklin FC, et al. (2013) Meiotic adaptation to genome duplication in *Arabidopsis arenosa*. *Curr Biol* 23(21):2151-6.
17. Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, et al. (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449(7161):463-7.
18. Tang H, Bowers JE, Wang X, Ming R, Alam M, Paterson AH (2008) Synteny and collinearity in plant genomes. *Science* 320(5875):486-8.
19. Ming R, VanBuren R, Wai CM, Tang H, Schatz MC, Bowers JE, et al. (2015) The pineapple genome and the evolution of CAM photosynthesis. *Nat Genet* 47(12):1435-42.

20. Aguilera A, Gomez-Gonzalez B (2008) Genome instability: a mechanistic view of its causes and consequences. *Nat Rev Genet* 9(3):204-17.
21. Karl R, Koch MA (2013) A world-wide perspective on crucifer speciation and evolution: phylogenetics, biogeography and trait evolution in tribe Arabideae. *Ann Bot* 112(6):983-1001.
22. Lysak MA, Koch MA, Beaulieu JM, Meister A, Leitch IJ (2009) The dynamic ups and downs of genome size evolution in Brassicaceae. *Mol Biol Evol* 26(1):85-98.
23. Ziolkowski PA, Blanc G, Sadowski J (2003) Structural divergence of chromosomal segments that arose from successive duplication events in the *Arabidopsis* genome. *Nucleic Acids Res* 31(4):1339-50.
24. Lysak MA, Berr A, Pecinka A, Schmidt R, McBreen K, Schubert I (2006) Mechanisms of chromosome number reduction in *Arabidopsis thaliana* and related Brassicaceae species. *Proc Natl Acad Sci USA* 103(13):5224-9.
25. Oliver KR, McComb JA, Greene WK (2013) Transposable elements: powerful contributors to angiosperm evolution and diversity. *Genome Biol Evol* 5(10):1886-901.
26. Lisch D (2013) How important are transposons for plant evolution? *Nature Reviews Genetics* 14(1):49-61.
27. Anderson LK, Lai A, Stack SM, Rizzon C, Gaut BS (2006) Uneven distribution of expressed sequence tag loci on maize pachytene chromosomes. *Genome Res* 16(1):115-22.
28. Fisher RA. *The Genetical Theory of Natural Selection*. Oxford: Oxford University Press; 1930. 318 p.
29. Hill WG, Robertson A (1966) The effect of linkage on limits to artificial selection. *Genet Res* 8(3):269-94.
30. Felsenstein J (1974) The evolutionary advantage of recombination. *Genetics* 78(2):737-56.
31. Bowers JE, Arias MA, Asher R, Avise JA, Ball RT, Brewer GA, et al. (2005) Comparative physical mapping links conservation of microsynteny to chromosome structure and recombination in grasses. *Proc Natl Acad Sci U S A* 102(37):13206-11.
32. Ma J, Bennetzen JL (2006) Recombination, rearrangement, reshuffling, and divergence in a centromeric region of rice. *Proc Natl Acad Sci U S A* 103(2):383-8.
33. Haddrill PR, Halligan DL, Tomaras D, Charlesworth B (2007) Reduced efficacy of selection in regions of the *Drosophila* genome that lack crossing over. *Genome Biol* 8(2):R18.
34. Hussin JG, Hodgkinson A, Idaghdour Y, Grenier JC, Goulet JP, Gbeha E, et al. (2015) Recombination affects accumulation of damaging and disease-associated mutations in human populations. *Nat Genet* 47(4):400-4.
35. Zhang X, Yazaki J, Sundaresan A, Cokus S, Chan SW, Chen H, et al. (2006) Genome-wide high-resolution mapping and functional analysis of DNA methylation in *Arabidopsis*. *Cell* 126(6):1189-201.
36. Zilberman D, Gehring M, Tran RK, Ballinger T, Henikoff S (2007) Genome-wide analysis of *Arabidopsis thaliana* DNA methylation uncovers an interdependence between methylation and transcription. *Nat Genet* 39(1):61-9.

37. Rabinowicz PD, Schutz K, Dedhia N, Yordan C, Parnell LD, Stein L, et al. (1999) Differential methylation of genes and retrotransposons facilitates shotgun sequencing of the maize genome. *Nat Genet* 23(3):305-8.
38. Regulski M, Lu Z, Kendall J, Donoghue MT, Reinders J, Llaca V, et al. (2013) The maize methylome influences mRNA splice sites and reveals widespread paramutation-like switches guided by small RNA. *Genome Res* 23(10):1651-62.
39. Li X, Zhu J, Hu F, Ge S, Ye M, Xiang H, et al. (2012) Single-base resolution maps of cultivated and wild rice methylomes and regulatory roles of DNA methylation in plant gene expression. *BMC Genomics* 13:300.
40. Green B, Bouchier C, Fairhead C, Craig NL, Cormack BP (2012) Insertion site preference of Mu, Tn5, and Tn7 transposons. *Mob DNA* 3(1):3.
41. Jiang N, Ferguson AA, Slotkin RK, Lisch D (2011) Pack-Mutator-like transposable elements (Pack-MULEs) induce directional modification of genes through biased insertion and DNA acquisition. *Proc Natl Acad Sci U S A* 108(4):1537-42.
42. Naito K, Zhang F, Tsukiyama T, Saito H, Hancock CN, Richardson AO, et al. (2009) Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. *Nature* 461(7267):1130-4.
43. Devos KM, Brown JK, Bennetzen JL (2002) Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. *Genome Res* 12(7):1075-9.
44. Bennetzen JL, Ma J, Devos KM (2005) Mechanisms of recent genome size variation in flowering plants. *Ann Bot* 95(1):127-32.
45. Tian Z, Rizzon C, Du J, Zhu L, Bennetzen JL, Jackson SA, et al. (2009) Do genetic recombination and gene density shape the pattern of DNA elimination in rice long terminal repeat retrotransposons? *Genome Res* 19(12):2221-30.
46. Hu TT, Pattyn P, Bakker EG, Cao J, Cheng JF, Clark RM, et al. (2011) The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nat Genet* 43(5):476-81.
47. SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL (1998) The paleontology of intergene retrotransposons of maize. *Nat Genet* 20(1):43-5.
48. Hollister JD, Gaut BS (2009) Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. *Genome Res* 19(8):1419-28.
49. Law JA, Jacobsen SE (2010) Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat Rev Genet* 11(3):204-20.
50. Matzke MA, Mosher RA (2014) RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. *Nat Rev Genet* 15(6):394-408.
51. Hollister JD, Smith LM, Guo YL, Ott F, Weigel D, Gaut BS (2011) Transposable elements and small RNAs contribute to gene expression divergence between *Arabidopsis thaliana* and *Arabidopsis lyrata*. *Proc Natl Acad Sci U S A* 108(6):2322-7.
52. Wang X, Weigel D, Smith LM (2013) Transposon variants and their effects on gene expression in *Arabidopsis*. *PLoS Genet* 9(2):e1003255.

53. Charlesworth B (2009) Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. *Nat Rev Genet* 10(3):195-205.
54. Lynch M, Conery JS (2003) The origins of genome complexity. *Science* 302(5649):1401-4.
55. Lynch M. *The Origins of Genome Architecture*. Sunderland, MA: Sinauer Associates; 2007. 389 p.
56. Lynch M (2007) The frailty of adaptive hypotheses for the origins of organismal complexity. *Proc Natl Acad Sci U S A* 104 Suppl 1:8597-604.
57. Koonin EV, Wolf YI (2010) Constraints and plasticity in genome and molecular-phenome evolution. *Nat Rev Genet* 11(7):487-98.
58. Koonin EV (2009) Evolution of genome architecture. *Int J Biochem Cell Biol* 41(2):298-306.
59. Roy SW, Gilbert W (2006) The evolution of spliceosomal introns: patterns, puzzles and progress. *Nat Rev Genet* 7(3):211-21.
60. Mitchell-Olds T, Willis JH, Goldstein DB (2007) Which evolutionary processes influence natural genetic variation for phenotypic traits? *Nat Rev Genet* 8(11):845-56.
61. Sawyer SA, Parsch J, Zhang Z, Hartl DL (2007) Prevalence of positive selection among nearly neutral amino acid replacements in *Drosophila*. *Proc Natl Acad Sci U S A* 104(16):6504-10.
62. Shapiro JA, Huang W, Zhang C, Hubisz MJ, Lu J, Turissini DA, et al. (2007) Adaptive genic evolution in the *Drosophila* genomes. *Proc Natl Acad Sci U S A* 104(7):2271-6.
63. Williamson SH, Hubisz MJ, Clark AG, Payseur BA, Bustamante CD, Nielsen R (2007) Localizing recent adaptive evolution in the human genome. *PLoS Genet* 3(6):e90.
64. Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, et al. (2006) Positive natural selection in the human lineage. *Science* 312(5780):1614-20.
65. Presgraves DC (2010) The molecular evolutionary basis of species formation. *Nat Rev Genet* 11(3):175-80.
66. Schluter D (2009) Evidence for ecological speciation and its alternative. *Science* 323(5915):737-41.
67. Bikard D, Patel D, Le Mette C, Giorgi V, Camilleri C, Bennett MJ, et al. (2009) Divergent evolution of duplicate genes leads to genetic incompatibilities within *A. thaliana*. *Science* 323(5914):623-6.
68. Masly JP, Jones CD, Noor MA, Locke J, Orr HA (2006) Gene transposition as a cause of hybrid sterility in *Drosophila*. *Science* 313(5792):1448-50.
69. Durand S, Bouche N, Perez Strand E, Loudet O, Camilleri C (2012) Rapid establishment of genetic incompatibility through natural epigenetic variation. *Curr Biol* 22(4):326-31.
70. Baker RJ, Bickham JW (1986) Speciation by monobrachial centric fusions. *Proc Natl Acad Sci U S A* 83(21):8245-8.
71. Rieseberg LH, Willis JH (2007) Plant speciation. *Science* 317(5840):910-4.
72. Jiang H, Guan W, Pinney D, Wang W, Gu Z (2008) Relaxation of yeast mitochondrial functions after whole-genome duplication. *Genome Res* 18(9):1466-71.

73. Costanzo MC, Bonnefoy N, Williams EH, Clark-Walker GD, Fox TD (2000) Highly diverged homologs of *Saccharomyces cerevisiae* mitochondrial mRNA-specific translational activators have orthologous functions in other budding yeasts. *Genetics* 154(3):999-1012.
74. Ellison CK, Niehuis O, Gadau J (2008) Hybrid breakdown and mitochondrial dysfunction in hybrids of *Nasonia* parasitoid wasps. *J Evol Biol* 21(6):1844-51.
75. Bomblies K, Lempe J, Epple P, Warthmann N, Lanz C, Dangl JL, et al. (2007) Autoimmune response as a mechanism for a Dobzhansky-Muller-type incompatibility syndrome in plants. *PLoS Biol* 5(9):e236.
76. Bakker EG, Toomajian C, Kreitman M, Bergelson J (2006) A genome-wide survey of *R* gene polymorphisms in *Arabidopsis*. *Plant Cell* 18(8):1803-18.
77. Chae E, Bomblies K, Kim ST, Karelina D, Zaidem M, Ossowski S, et al. (2014) Species-wide genetic incompatibility analysis identifies immune genes as hot spots of deleterious epistasis. *Cell* 159(6):1341-51.
78. Sawamura K, Yamamoto MT (1997) Characterization of a reproductive isolation gene, zygotic hybrid rescue, of *Drosophila melanogaster* by using minichromosomes. *Heredity* 79:97-103.
79. Ferree PM, Barbash DA (2009) Species-Specific Heterochromatin Prevents Mitotic Chromosome Segregation to Cause Hybrid Lethality in *Drosophila*. *PLoS Biol* 7(10).
80. Charlesworth D, Willis JH (2009) The genetics of inbreeding depression. *Nat Rev Genet* 10(11):783-96.
81. Davenport CB (1908) Degeneration, albinism and inbreeding. *Science* 28(718):454-5.
82. Bruce AB (1910) The Mendelian theory of heredity and the augmentation of vigor. *Science* 32(827):627-8.
83. Jones DF (1917) Dominance of linked factors as a means of accounting for heterosis. *Proc Natl Acad Sci U S A* 3:310-2.
84. Shull GH (1908) The composition of a field of maize. *J Hered os*-4(1):296-301.
85. East EM (1908) Inbreeding in corn. *Rep Conn Agric Exp Stn* 1907:419-28.
86. Beilstein MA, Nagalingum NS, Clements MD, Manchester SR, Mathews S (2010) Dated molecular phylogenies indicate a Miocene origin for *Arabidopsis thaliana*. *Proc Natl Acad Sci U S A* 107(43):18724-8.
87. Kawabe A, Hansson B, Hagenblad J, Forrest A, Charlesworth D (2006) Centromere locations and associated chromosome rearrangements in *Arabidopsis lyrata* and *A. thaliana*. *Genetics* 173:1613-9.
88. Morgan HD, Sutherland HG, Martin DI, Whitelaw E (1999) Epigenetic inheritance at the agouti locus in the mouse. *Nat Genet* 23(3):314-8.
89. Martin A, Troadec C, Boualem A, Rajab M, Fernandez R, Morin H, et al. (2009) A transposon-induced epigenetic change leads to sex determination in melon. *Nature* 461(7267):1135-8.
90. Das OP, Messing J (1994) Variegated phenotype and developmental methylation changes of a maize allele originating from epimutation. *Genetics* 136(3):1121-41.
91. Liu J, He Y, Amasino R, Chen X (2004) siRNAs targeting an intronic transposon in the regulation of natural flowering behavior in *Arabidopsis*. *Genes Dev* 18(23):2873-8.

92. Studer A, Zhao Q, Ross-Ibarra J, Doebley J (2011) Identification of a functional transposon insertion in the maize domestication gene *tb1*. *Nat Genet* 43(11):1160-3.
93. Romero IG, Ruvinsky I, Gilad Y (2012) Comparative studies of gene expression and the evolution of gene regulation. *Nat Rev Genet* 13(7):505-16.
94. Takuno S, Gaut BS (2013) Gene body methylation is conserved between plant orthologs and is of evolutionary consequence. *Proc Natl Acad Sci U S A* 110(5):1797-802.
95. Takuno S, Ran J-H, Gaut BS (2016) Evolutionary patterns of genic DNA methylation vary across land plants. *Nature Plants* 2(2):15222.
96. Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, et al. (2008) Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. *Nature* 452(7184):215-9.
97. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, et al. (2008) Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* 133(3):523-36.
98. Chodavarapu RK, Feng S, Bernatavichute YV, Chen PY, Stroud H, Yu Y, et al. (2010) Relationship between nucleosome positioning and DNA methylation. *Nature* 466(7304):388-92.
99. Baubec T, Finke A, Scheid OM, Pecinka A (2014) Meristem-specific expression of epigenetic regulators safeguards transposon silencing in *Arabidopsis*. *EMBO Rep* 15(4):446-52.
100. Turelli M (1984) Heritable Genetic-Variation Via Mutation Selection Balance - Lerch Zeta Meets the Abdominal Bristle. *Theor Popul Biol* 25(2):138-93.
101. Lyttle TW (1991) Segregation distorters. *Annu Rev Genet* 25:511-57.
102. Corbett-Detig RB, Zhou J, Clark AG, Hartl DL, Ayroles JF (2013) Genetic incompatibilities are widespread within species. *Nature* 504(7478):135-7.
103. Crow JF (1993) Mutation, mean fitness, and genetic load. *Oxford Surv Evol Biol* 9:3-42.
104. Fishman L, Saunders A (2008) Centromere-associated female meiotic drive entails male fitness costs in monkeyflowers. *Science* 322(5907):1559-62.

9. Acknowledgements

I am immensely grateful to Detlef Weigel for creating a wonderful scientific environment that continually fostered intellectual growth, independence, and creativity in each of its members. Thanks, Detlef, for being an outstanding mentor and for sharing your infectious enthusiasm for science. A huge thank you to Rebecca Schwab for ensuring that we can all do great science. I'm appreciative of all of the Weigelworld members for making Tübingen a wonderful place to spend nearly five years of my life, and in particular to Ignacio Rubio Somoza, Hernán Burbano, Claude Becker, George Wang, Beth Rowan, Pablo Manavella, Natalia Camomilli, Carmen Martín Pizarro, David Posé Padilla, Eunyoung Chae, Jörg Hagmann, Tobias Langenecker, Bonnie Fraser, Cameron Weadick, and Mike and Tara Christie. I'm particularly grateful to my parents for teaching me the value of education and for supporting me throughout the entire educational process, no matter how far away it took me from them. Finally, I'd like to acknowledge Daniel Koenig for his scientific mentorship, emotional support, and for sharing in a wonderful German adventure.

Danelle Seymour
Max Planck Institute for Developmental Biology
Department of Molecular Biology
Spemannstrasse 37-39
72076 Tübingen, Germany
danelle.seymour@tuebingen.mpg.de

Professional Preparation

University of California, Davis	Genetics	B.S., 2007
University of California, Davis	Genetics	M.S., 2011
Max Planck Institute for Developmental Biology	Biology	Ph.D., Expected 2016

Appointments

Max Planck Institute for Developmental Biology, Tübingen, Germany Graduate Student Researcher; Advisor: Detlef Weigel	2011 - Current
University of California, Davis, USA Graduate Student Researcher; Advisor: Julin Maloof	2010 - 2011
Monsanto Company (Vegetable Division), Woodland, USA Research Associate	2007 - 2010
University of California, Davis, USA Undergraduate Researcher, Michelmore Laboratory	2006 - 2007

Publications

- Karlsson P, Christie MD, Seymour DK, Wang H, Wang X, Hagmann J, Kulcheski F, Manavella P (2015). KH domain protein RCF3 is a tissue-biased regulator of the plant miRNA biogenesis cofactor HYL1. *Proc Natl Acad Sci USA* 110:12120-12125.
- Rawat V, Abdelsamad A, Pietzenuk B, Seymour DK, Koenig D, Weigel D, Pecinka A, Schneeberger K (2015). Improving the annotation of *Arabidopsis lyrata* using RNA-seq data. *PLoS One* 10:e0137391.
- Seymour DK*, Koenig D*, Hagmann J, Becker C, Weigel D (2014). Evolution of DNA methylation patterns in the Brassicaceae is driven by differences in genome organization. *PLoS Genet* 10:e1004785.
- Rugnone ML, Faigón Soverna A, Sanchez SE, Schlaen RG, Hernando CE, Seymour DK, Mancini E, Chernomoretz A, Weigel D, Más P, Yanovsky MJ (2013). LNK genes integrate light and clock signaling networks at the core of the *Arabidopsis* oscillator. *Proc Natl Acad Sci USA* 110:12120-12125.
- Seymour DK, Filiault DL, Henry IM, Monson-Miller J, Ravi M, Pang A, Comai L, Chan SW, Maloof JN (2012). Rapid creation of *Arabidopsis* doubled haploid lines for quantitative trait locus mapping. *Proc Natl Acad Sci USA* 109:4227-4232.
- Sandlin K, Prothro J, Heesacker A, Khalilian N, Okashah R, Ziang W, Bachlava E, Caldwell DG, Taylor CA, Seymour DK, White V, Chan E, White C, Safran D, Graham E, Knapp S, McGregor C (2012). Comparative mapping in watermelon [*Citrullus lanatus* (Thunb.) Matsum. et Nakai]. *Theor Appl Genet* 125:1603-1618.

*These authors contributed equally

Submitted Manuscripts

- Seymour DK*, Chae E*, Grimm D*, Martín-Pizarro C, Vasseur F, Rakitsch B, Borgwardt K, Koenig D, Weigel D. The genetic architecture of non-additive hybrid phenotypes in *Arabidopsis thaliana*. (In revision).
- Rowan BA, Seymour DK, Chae E, Lundberg D, Weigel D. Methods for genotyping-by-sequencing. *Methods in Molecular Biology* (In revision).

*These authors contributed equally

Oral Presentations

"Exploring the genetics and genomics of <i>Arabidopsis thaliana</i> and its relatives" IAL CONICET – Santa Fe, Argentina	2015
"Exploring the genetics and genomics of <i>Arabidopsis thaliana</i> and its relatives" Postdoctoral interview at UC Irvine - Irvine, USA	2015
"Exploring the genetics and genomics of <i>Arabidopsis thaliana</i> and its relatives" Postdoctoral interview at UCLA - Los Angeles, USA	2015
"The evolution of DNA methylation patterns in the <i>Brassicaceae</i> " 2014 Ph.D. Symposium - Tübingen, Germany	2014
"DNA methylome comparisons across three <i>Brassicaceae</i> species" 7th RegioPlantScience Meeting - Stuttgart, Germany	2014
"DNA methylome comparisons in three <i>Brassicaceae</i> species" StEvE (Students in Evolution and Ecology) - Tübingen, Germany	2012
"DNA methylome comparisons in three <i>Brassicaceae</i> species" 10th International Plant Molecular Biology Congress - Jeju, South Korea	2012

Poster Presentations

"A species-wide screen for segregation distortion in <i>Arabidopsis thaliana</i> " EMBO Workshop - Åkersberga, Sweden	2015
"The genetic basis of dominance in <i>Arabidopsis thaliana</i> " Ph.D. Student Retreat - Weil der Stadt, Germany	2014
"DNA methylome comparisons in three <i>Brassicaceae</i> species" EMBO EMBL Symposium - Heidelberg, Germany	2013
"QTL Mapping in an <i>Arabidopsis</i> Doubled Haploid Population" Plant Cell Biology Retreat - Asilomar, USA	2010
"High Throughput Vegetable SNP Discovery" Above and Beyond Awards - St. Louis, USA	2010

Teaching Experience

Introduction to statistics and the computing package R 3 session course on statistics, experimental design, and R programming.	2013, 2014
Introduction to the R programming language Weeklong training course for an NSF funded undergraduate internship.	2011

Service

Reviewer for Molecular Ecology and Genome Biology and Evolution	2015, 2016
Max Planck Institute Ph.D. Representative	2013 - 2015
Max Planck Institute Ph.D. retreat (Primary organizer, 4 invited speakers)	2014

Awards and Fellowships

NSF Postdoctoral Research Fellow in Biology	2016
Life Sciences Postdoctoral Fellowship (finalist)	2016
Max Planck Institute Fellowship	2011 - 2015
Above-and-Beyond, Monsanto	2009