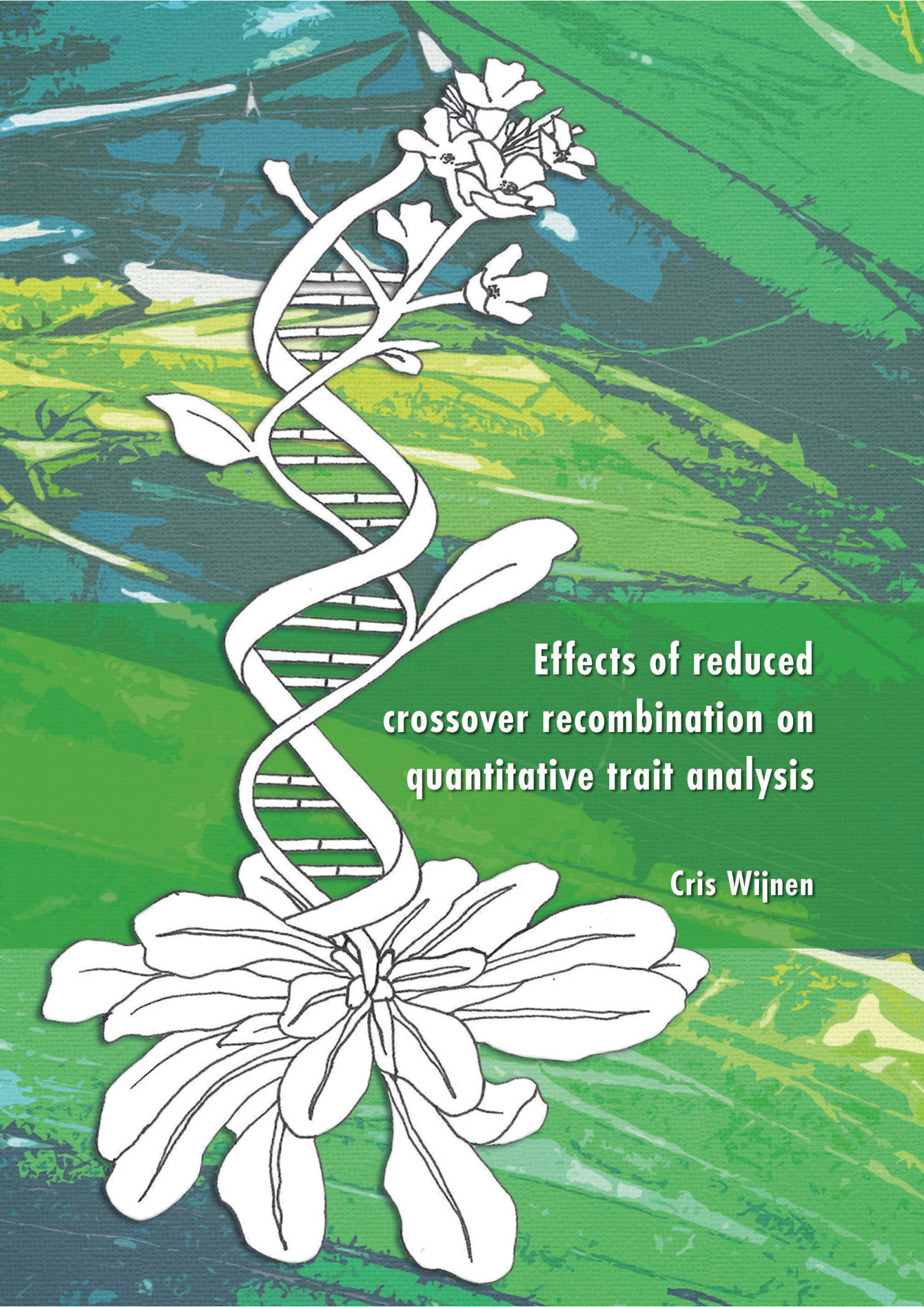


Effects of reduced crossover recombination on quantitative trait analysis

Cris Wijnen

Propositions

1. Even in a post-high-throughput-genotyping era, chromosome substitution lines are a valuable complement to the array of current mapping resources.
   (this thesis)

2. For the initial detection of genetic effects, crossover recombination is undesirable as it needlessly complicates the analyses.
   (this thesis)

3. The European hamster -Korenwolf- is a better representative in relation to agricultural biodiversity than the honey bee, due to its cuddliness, and tendency for cannibalism when fed with monoculture maize diets (Tissier et al., Proc Biol Sci, 284, 2017).

4. Proper data management and content curation is the most important aspect in a future with self-learning algorithms.

5. Endurance sports are an effective way to balance a scientists mental challenges.

6. Due to digitalization costumers are carrying out their own costumer-care, but personal contact can never be replaced.

7. Without additional regulation, pollution of cycling paths with motorized vehicles is a hazard to the healthy Dutch cycling culture.

Propositions belonging to the thesis, entitled

Effects of reduced crossover recombination on quantitative trait analysis

Cristian L. Wijnen
Wageningen, 14 October 2019

Effects of reduced crossover recombination on quantitative trait analysis

Cris L. Wijnen

Thesis committee

Promotors
Dr J.J.B. Keurentjes
Associate professor, Laboratory of Genetics
Wageningen University & Research

Prof. Dr F.A. van Eeuwijk
Professor of Applied Statistics
Wageningen University & Research

Co-promotors
Dr T.G. Wijnker
Laboratory of Genetics
Wageningen University & Research

Dr M.P. Boer
Researcher, Biometris
Wageningen University & Research

Other members
Prof. Dr J.E. Kammenga, Wageningen University & Research
Dr C. Lelivelt, Rijk Zwaan Breeding B.V., Fijnaart
Dr J.M. Jiménez-Gómez, INRA Centre de Versailles-Grignon, France
Prof. D. Geelen, Ghent University, Belgium

This research was conducted under the auspieces of the Graduate School Experimental Plant Sciences (EPS)

Effects of reduced crossover recombination on quantitative trait analysis

Cristian Lucas Wijnen

Thesis
submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus,
Prof. Dr A. P. J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public on Monday October 14 2019
at 1.30 p.m. in the Aula

Cris L. Wijnen
Effects of reduced crossover recombination on quantitative trait analysis,
176 pages.

PhD thesis, Wageningen University, Wageningen, The Netherlands (2019)
With references, with summary in English

ISBN: 978-94-6395-120-3
DOI: https://doi.org/10.18174/500002

TABLE OF CONTENTS

"I only went out for a walk and finally concluded to stay out till sundown, for going out, I found, was really going in."

John Muir

Chapter 1

General introduction

[Adapted from Wijnen and Keurentjes. Curr Opin Plant Biol. 18: 103-109. 2014.]

One of the striking observations in the plant kingdom is the vast amount of standing natural variation in quantitative traits displayed within and between species (*1*). Initially, human cultures of gatherers searched among these plants for those with interesting qualities and collected the edible parts for consumption. About 10,000 years ago people took notion of the heredity of these evolutionary shaped traits and started to exploit this phenotypic diversity as farmers to improve their crops. Since then, without much notion of the genetic mechanisms, crop varieties changed gradually towards accommodating the needs of human society by domestication.

During the last two centuries the biological scientific community described how selection and inheritance affect evolutionary and domestication processes. Today, modern science takes advantage of the diversity in quantitative traits as a means to elucidate the genetic regulation of biological processes and the evolutionary forces driving selection (*2*). For each of these purposes it is pivotal to identify the genetic loci, or quantitative trait loci (QTL), that are causal for the observed phenotypic variation. Typically a search for such QTLs is initiated by creating genetic mapping resources in which genetic variation between parental lines segregates in recombinant offspring. QTLs are detected when specific chromosome segments co-segregate with specific traits. As will be detailed later in this chapter, the segregation of chromosome segments in offspring can take different forms depending on the offspring type. But essentially, segregating segments range from small segments of chromosomes to entire chromosomes. Mapping populations that differ in this aspect are referred to as having respectively a complex or simple genetic architecture. The statistical model to estimate genetic effects is limited for genetically simple mapping populations, but the creation is not easy as will become clear later.

This thesis addresses the question as to whether genetically simple mapping populations in *Arabidopsis thaliana* based to a large extent on segregating chromosomes may in fact improve the detection of QTLs in offspring. It is hypothesized that less resolution would lead to more power, and this could be used to identify even interaction effects in such populations. In this first chapter, the general basics of QTL mapping in plant mapping populations are explained, followed by an overview of recent developments that caused a renewed interest in more simple population types. This chapter finishes by an overview that introduces the subsequent chapters that provide insight in the use of such simple populations for genetic mapping.

Recombination creates new genotypes

In genetic mapping the intent is to look for associations between the manifestation of traits and genetic loci that segregate in offspring (*3*). The segregation of alleles in gametes of the offspring of a cross between two distinct founder lines results from meiotic recombination. Meiosis is the process in which a somatic cell halves its chromosome number in two subsequent cell divisions to give rise to four haploid gametes. While meiosis proceeds along a strictly regulated process divided into separate phases, one important event that occurs during the first meiotic division is meiotic recombination (*4*).

Meiotic recombination results in the redistribution of the parental genetic information over reproductive cells. Recombination is generally divided into two different processes: crossover recombination and random chromosome assortment, both of which take place during the first meiotic division. For the formation of crossovers, stretches of DNA are exchanged between non-sister chromatids creating new allele combinations within a chromosome. In *Arabidopsis thaliana* about one to two crossover events occur per chromosome pair, which is mainly influenced by the length of a chromosome (*5*). Crossover recombination thus creates new non-allelic combinations that could contribute to new phenotypic variation.

In eukaryote genomes, containing multiple independent chromosomes, crossover recombination has a second essential function in ensuring that homologous chromosomes become physically attached for proper disjunction (*4*). Because of this physical attachment during the first meiotic division, homologous chromosomes can segregate to opposite poles independently of other pairs. Since not all homologs of the same founder genotype are directed to the same pole but segregate randomly, this is the second process that generates new allelic combinations in the resulting gametes. In the subsequent second meiotic division sister-chromatids separate to give rise to haploid spores. While haploid gametes of homozygous genotypes undergo this process with no consequence to the resulting genotype, haploid gametes of heterozygotes represent a mosaic of the two parental genomes due to the independent assortment of, and recombination between homologous chromosomes.

Linking genotype to phenotype

As a consequence of meiotic recombination, all gametes of an F_1 hybrid (the result of a cross of two homozygous founder lines) are unique, composed of a specific genetic combination of the two founder lines. When such gametes are used to produce a segregating mapping population, phenotypic variation between individuals can be

associated to the descent of a specific genomic region. Recombination then not only allows the genetic study of variation in properties of the parental lines, but also of the effects of new allelic combinations not previously present in the founder genotypes. Offspring can thus be classified in genotypic classes according to their genotype at single or multiple loci together (*6*).

By performing controlled experiments, for each offspring genotype an estimate for its corresponding phenotypic value of a particular trait is obtained. Association between phenotypic value and genomic region not only depends on precise measurement and estimation of phenotypic values for individual genotypes, but also on the number of offspring in the investigated population; with increasing population size, a more precise estimate of the genotype effect on a phenotype can be obtained (*3*). Indeed, it is not unusual to obtain mapping populations of hundreds to thousands of individual genotypes, which translates in sufficient mapping power and resolution (*7-9*). To establish whether there is a significant mean difference in phenotypic value between genotypes, statistical methods are used. Essentially, mapping is the assessment of a significant association between the genotype at one, or multiple, specific DNA marker position(s) and the corresponding estimates for a phenotype (**Fig. 1**).

From a single marker to high-density maps

The use of multiple markers and large populations of segregating lines has enabled the construction of linkage maps (*10*), that are crucial when associating genomic loci with phenotypic traits and QTLs. In such maps it is visualised how much recombination, on average, takes place between markers on a chromosome. Genetic linkage maps are extremely useful if no genomic information of physical marker positions are available, as was the case in the early days of genetic mapping in model species and nowadays still is for many crops and wild species.

A strong association of a marker, even without any positional information, with trait variation is in most cases sufficient for introgression breeding and marker-assisted-selection (MAS) purposes (*3*). However, emerging sequencing technologies over the last decade have aided in increasing the total number of markers that can be analysed, creating high-density maps that approximate gene density resolution. Genome-wide sequence information can also assist in estimating the effects of single nucleotide polymorphisms (SNPs), which is extremely helpful for the fundamental understanding of gene function (*11*).

The statistical methods to detect QTL advanced accordingly. Initially, single-marker analyses were performed with a *Students t-tests* (for two genotypic classes at a single locus; AA vs. aa) or *Analysis of Variance* (ANOVA; for two or three genotypic classes at a single locus; AA, Aa and aa) which only considered the markers under study (*3*). Equivalently a regression of the phenotype can be performed on the number of A alleles (0, 1 or 2).

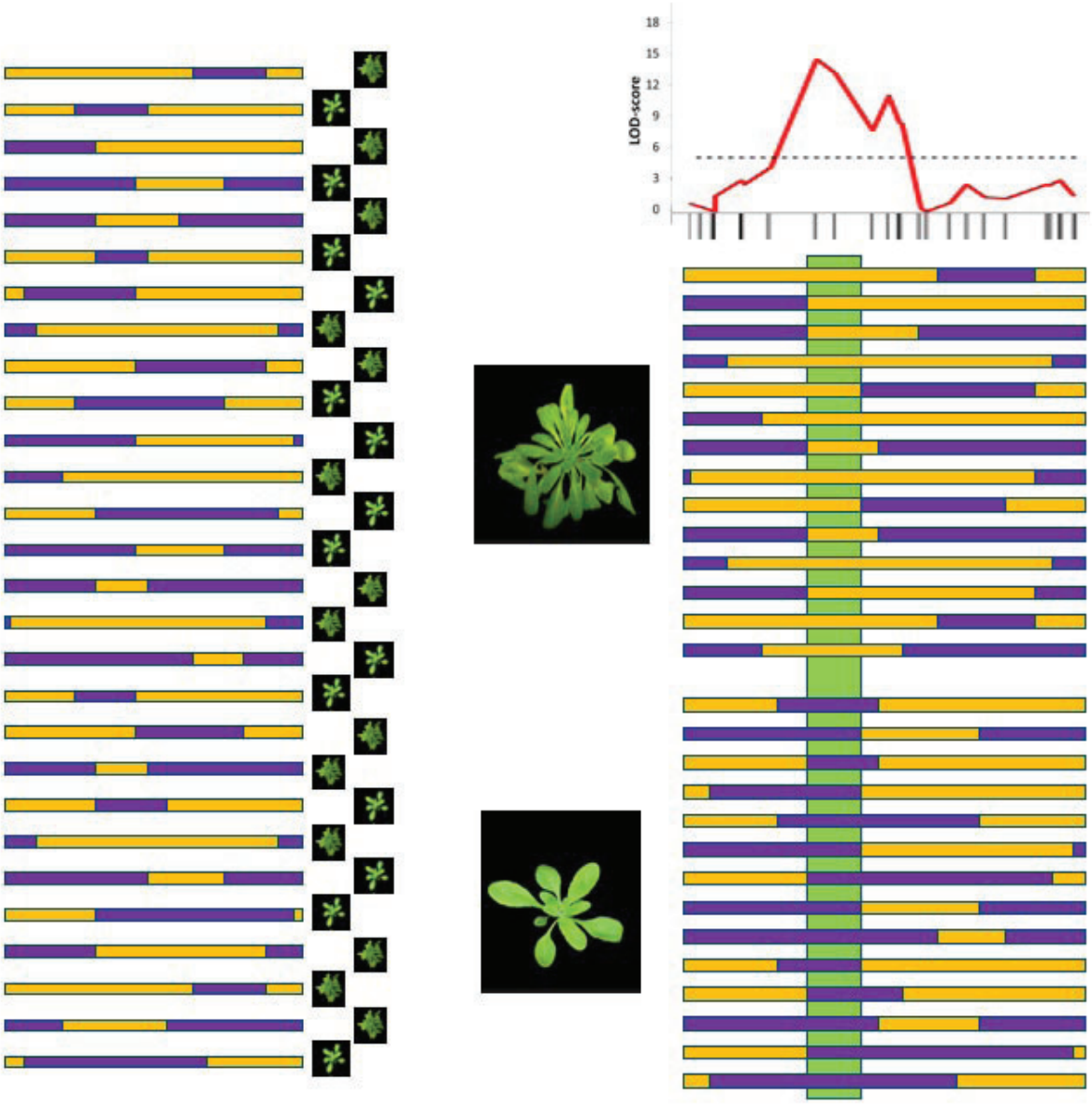


FIGURE 1 | A schematic overview of QTL mapping. In this oversimplified view of QTL mapping the three most important aspects for identifying QTLs are presented, that is genotyping, phenotyping and statistical analysis. In the left panel a single homozygous chromosome is shown for genotypically different individuals, the colours indicate a different parental origin of genomic regions. Next to each genotype, the phenotype is presented as an Arabidopsis rosette with quantitative variation in leaf number. After classifying each individual based on a marker's genotype (green box) on the right panel, a statistical analysis tests for association between the observed phenotypic variation and genotypic variation at this locus, The test statistic for the association is represented by the profile plot on the top.

Such a regression approach is also used in simple interval mapping (SIM), although here additional hypothetical marker genotypes are calculated for positions in between observed markers, based on the flanking marker genotypes and the genetic distance between the observed markers and the positions of the hypothetical markers (*12*). For these hypothetical marker positions, QTL mapping is performed using regression of the phenotype on the expected number of A alleles (between 0 and 2).

This approach was further developed into composite interval mapping (CIM), where more power to detect secondary QTLs is obtained by absorbing the variation introduced by segregating QTLs elsewhere in the genome. This is done through inclusion of markers associated with detected QTLs as cofactors during a next genome scan. By doing so, the effects of these QTLs are removed from the error term of the statistical model. Further generalization of QTL mapping for multiple traits and/or environments requires the use of mixed models. Although the statistical methodology of detecting QTLs has advanced enormously, the essence of genetic mapping still lies in the detection of significant associations between genotypic and phenotypic variation (*13*).

Different genetic mapping resources for detection of QTLs

The ability to identify genes underlying specific traits has always been of major interest to the scientific community. The focus of QTL mapping in humans is on disease variant detection, in plant breeding it is towards crop improvement, while in model systems like Arabidopsis the elucidation of more fundamental processes are of additional interest. As mentioned, the capacity to genotype individuals on a genome-wide scale in addition to modern ways of accurate phenotyping has been crucial in genetic screens. The impracticalities of genotyping protocols and the large sizes of mapping populations that are needed to acquire sufficient genotypic replication and hence statistical power led to high costs for genetic screens.

However, during the past decades multiple types of genetic resources have been developed, driven by technological advancements, and used to elucidate the genotype-phenotype relationship (*14*). These resources can be as diverse as collections of wild germplasm (*15*), experimentally derived populations (*16*), e.g. through crossing and backcrossing, and even assembled panels of artificially generated mutants or transgenic genotypes (*17*). Mapping resources constructed from crosses between inbred parents consisted initially of F_2 populations and backcross (BC) progeny. However, such (partially) heterogeneous genotypes segregate in future generations. With the exception for clonal species, this restricts the use of the acquired genotype to a single individual in each generation and it thus can be used only in a single experiment.

To circumvent this, many current mapping populations consist of homozygous lines in which the genotype is fixed and thus can be propagated indefinitely without changing the genetic make-up, creating so called immortal genotypes (*10, 18*). This facilitates the inclusion of genotype replications and their use in multiple different experiments or environments. Additionally, in such homozygous populations there is no interference of dominance effects. Although dominance effects could lead to more power for QTL detection, homozygous genotypes are usually preferred over heterozygous genotypes to make genetic analysis more straightforward.

To achieve complete homozygosity, however, typically eight to ten generations of inbreeding are necessary which has severely hampered the *ad hoc* development of populations for specific traits. Indeed, only a limited number of genotyped immortal populations such as near isogenic lines (NILs) and recombinant inbred lines (RILs) have so far been created in Arabidopsis (*19-21*), mainly because of the time investments needed. Therefore, many genetic studies have been performed on the same genetic resources, which has restricted the full use of the natural diversity present within the species. In **Chapter 2** a novel mapping resource is described that in part responds to the need of including additional allelic variation within genetic populations of Arabidopsis.

Innovations in population types and mapping approaches

Genotyping has become sufficiently easy to (re)consider alternative types of mapping populations, especially with the advent of next generation genotype-by-sequencing technology. This has inspired the use of larger heterozygous populations, such as F_2 populations, but also allowed the assembly of large panels of wild accessions for genome wide association studies (GWAS). In GWAS, exploiting panels of hundreds to thousands of wild accessions, a much wider genetic and assumingly phenotypic diversity of the species is investigated. Making use of the large number of recombination events that have historically shaped the genotype of each accession, such GWAS panels permit mapping with a resolution that approximates genome-wide coverage of genes by SNPs (*22*). With this impulse of large-scale genotyping possibilities, research questions can now be addressed that were previously impossible to answer using conventional approaches (*15, 23, 24*).

Although mapping resolution and allelic sampling are unprecedented, GWAS suffers from several drawbacks. The large population sizes, for example, hamper large-scale phenotyping. More importantly, confounding effects of population structure, the independent segregation of many contributing loci and the presence of multiple alleles per locus increases complexity and reduces mapping power accordingly (*25*).

To combine the main advantages of GWAS (more allelic variation and higher resolution present) and the strength of RILs (equal representation of haplotypes and high repeatability of genotype observations), attempts have been made to integrate both approaches (*26*). For instance, to increase the recombination frequency and, hence, the mapping resolution in experimental populations, multiple rounds of intermating can be applied before inbreeding to create advanced intercross RILs (*27, 28*). In addition, incorporating multiple parents in the crossing design will elevate the allelic diversity in the population. Such multi-parental populations instantaneously increase recombination frequency because of the necessary extra generations for intercrossing: F_1, F_2 and later generations are crossed to combine multiple founder genomes in a single progeny individual (*16, 29*).

Currently two of such multi-parental populations are available for Arabidopsis, viz. the multiparent advanced generation inter-cross (MAGIC) (*16*) and the Arabidopsis multiparent RIL (AMPRIL) population (*29*). Although these populations alleviate some of the shortcomings attributed to GWAS (e.g. population structure and influences of minor allele frequencies) and conventional populations (e.g. low resolution and low allelic diversity), a vast amount of time and effort is required for producing them. Multi-parental populations are useful, but it can also be argued that the complexity of such populations is needlessly elevated and actually the disadvantages of GWAS and biparental populations are combined. In this thesis the opposite is attempted. By decreasing the crossover recombination frequency a simpler population architecture is obtained that arguably provides unique opportunities for genetic mapping.

Generation of homozygous genotypes using doubled haploids

Although homozygous populations have certain advantages over heterozygous populations, in most species the many generations required to obtain homozygous populations is usually a problem. This can be circumvented by the production of homozygous lines directly from haploid gametes of an F_1 hybrid. The generation of haploid plants and thereafter doubled haploids (DHs) is the preferred method for the quick creation of a homozygous mapping resource in many crop species such as barley, maize and wheat (*30*).

Basically, a DH is produced after spontaneous chromosome doubling of a haploid, creating instantaneously a homozygous genotype (for a diploid species). DHs can be obtained through maternal (gynogenesis) or paternal (androgenesis) gametes using intraspecific and wide hybridization, different kinds of microspore or pollen treatments, or embryogenesis (*31*). A DH population with largely similar features as a RIL population can be developed (DHs have lower resolution due to the single meiotic recombination) in no more than three generations, instead of eight.

Arabidopsis thaliana has been a versatile model species for the mapping of quantitative traits since the early eighties of last century (*32*). The short generation times, high number of seed set, wide distribution range, capacity of outcrossing and tolerance to inbreeding has made it the species of choice for the development of many genetic mapping resources (*33*). Nonetheless, the species has remained recalcitrant to the development of several highly desired technologies. Notably, even though DHs can be produced in close relatives of Arabidopsis in the *Brassica spp*, haploid formation was unsuccessful until the development of a haploid inducer line (*31, 34, 35*).

Genome elimination for doubled haploid production in Arabidopsis

Genome elimination can be defined as the loss of a complete parental set of chromosomes from a zygote after hybridization. While this phenomenon was described earlier for interspecific crosses (*36-39*), the causal process inducing this usually randomly occurring event has been unknown for years. However, now it is known that genome elimination can be caused by modification of the *CENTROMERIC HISTONE H3/HISTONE H3-LIKE CENTROMERIC PROTEIN 12* (*40, 41*).

In their seminal paper of 2010, Ravi and Chan complemented such a *cenh3/htr12* mutant with a construct in which the histone tail of the CENH3/HTR12 was replaced by a GFP-tagged tail of histone 3 (H3), creating the so-called *GFP-tailswap* line. The complemented mutant *GFP-tailswap* is viable and sets seeds. When outcrossing *GFP-tailswap* with a wild-type line, initially a normal zygote is formed but during the first mitotic divisions of this zygote the genome derived from the *GFP-tailswap* parent is lost due to defective spindle attachment (**Fig. 2**) (*42*). This leads effectively to a haploid zygotic genome derived from only the wild-type parent gamete. Although the haploid plants are largely sterile (no homologs can pair and thus no stable bivalents can be formed), incidental somatic doubling or the formation of 2n gametes can give rise to homozygous DHs.

This *GFP-tailswap* or transgenic haploid inducer has greatly expanded the possibilities for development of new genetic resources in Arabidopsis (*43*). For instance it has been used for synthetic apomixes (*44*), and for other applications in genetics studies (*45*). Also the easy generation of DH populations has become feasible with the use of the *GFP-tailswap*. When the *GFP-tailswap* is fertilised with the pollen of an heterozygous F_1 plant all the haploid seeds formed contain zygotes that differ in their number and positions of recombinations. However, since the publication of the haploid inducer several years ago, only three DH populations have been reported for Arabidopsis (*46-48*). Although the first DH population was valuable in showing the potential of DHs as Arabidopsis genetic

mapping resources (*47*), it was especially the second DH population that showed the innovative possibilities of the DH methodology in Arabidopsis by combining it with complete suppression of crossover recombination (*46*). In the next section, it is explained how this was achieved and why this was important.

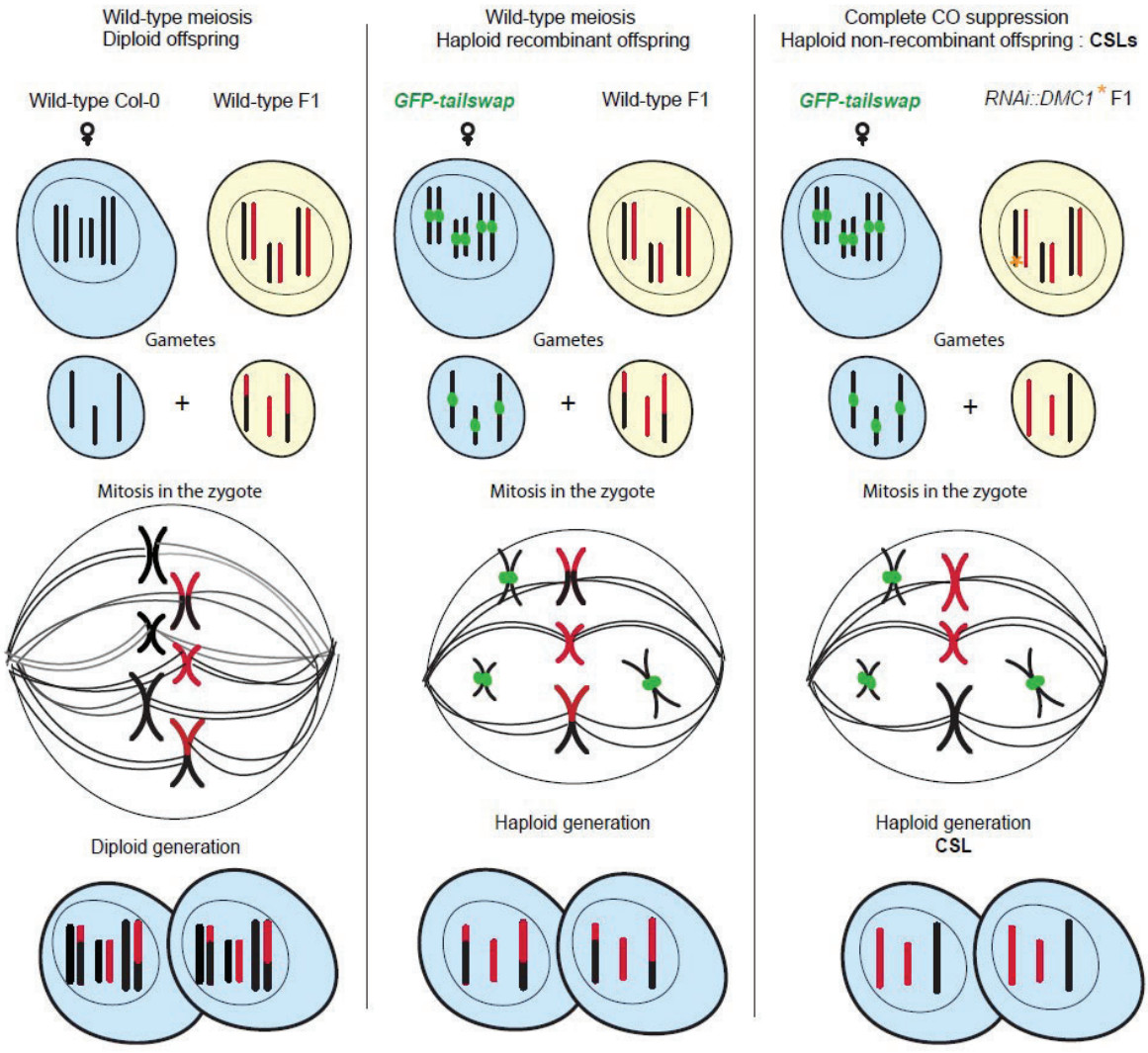


FIGURE 2 | Comparison of the different uses for the *GFP-tailswap* line as discussed in this thesis. In a standard wild-type backcross between an F_1 and an inbred parental line, meiosis and subsequent mitosis in the zygote leads to diploid offspring with recombined and non-recombinant chromosomes (left panel). For haploid production a wild-type F_1 is crossed with the *GFP-tailswap* line (middle panel). Due to the *GFP-tailswap* (depicted by green fluorescent tags at the centromeres) the *GFP-tailswap* line chromosomes are lost during the first couple of mitotic divisions of the zygote, resulting in haploid recombinant progeny. In reverse breeding (right panel), an F_1 transformed with a dominant negative *RNAi::DMC1* construct is used instead of a wild-type F_1 during haploid production (represented by the orange star), this results in non-crossover haploid offspring. Note that the resulting offspring always contains the cytoplasm of the maternal *GFP-tailswap*.

Genome elimination and suppression of crossovers allow "Reverse Breeding"

In a proof-of-principle study Wijnker et al. (*46*) provided practical evidence for the concept of reverse breeding (*49*). Instead of creating the hybrid from parental lines, in reverse breeding parental lines can be obtained from an hybrid. This is achieved by suppression of crossover recombination in the hybrid genotype to produce DHs that encompass only non-crossover chromosomes.

By use of a transgenic dominant negative RNAi construct targeting *DISRUPTED MEIOTIC cDNA 1* (*DMC1*; which is essential for the mediation of inter-homolog interactions during meiosis) meiotic recombination is altered such that crossover recombination does no longer take place (**Fig. 2**). This leads to the random segregation of the non-recombinant homologs during meiosis I, and the subsequent segregation of chromatids at meiosis II. Although crossover recombination has been abolished, random segregation of the founder chromosomes still occurs. Therefore, when executed in an F_1 hybrid plant, the gametes consist of different combinations of the chromosomes derived from the parents of the initial hybrid. These gametes can be grown into haploid offspring through crossing with *GFP-tailswap* (*50*). The haploid lines are subsequently grown and self-fertilised to produce homozygous diploid lines.

It should be noted that abolishing crossover recombination has the consequence of creating (semi-)sterile plants since homolog pairing is skipped and proper disjunction of the chromosomes relies on chance events. Additionally when the *GFP-tailswap* is used as the maternal plant, all haploid offspring contains the cytoplasm of the *GFP-tailswap* genotype, excluding a potential source of variation between the haploids.

The interesting aspect of these acquired genotypes is that they contain one or multiple intact chromosome(s) from one founder parent, while all other chromosomes descent from a second parent. Therefore, the more common name to refer to such genotypes has become chromosome substitution lines (CSLs; a.k.a. CSS for strains or even consomic strains in non-plant species). Such genotypes can form perfect complementary genotypes that can serve as parental lines to recreate the F_1 hybrid genotype, as was shown for reverse breeding.

The strategy of reverse breeding is elaborated in **Chapter 5** by providing a more efficient approach to produce reverse breeding offspring through reduced crossover recombination instead of complete absence. Besides the development of reverse breeding, the generation of DH resources derived from the *GFP-tailswap* line has opened new trajectories for the study of natural variation in Arabidopsis, and a number of these possibilities are pursued within this thesis.

The formation of non-crossover chromatids

F_1 hybrid offspring genotypes containing only non-crossover chromosomes, such as produced by reverse breeding, have themselves been considered as a potentially powerful mapping resource. It is possible to generate such genotypes also through classical genetic approaches, due to the occasional occurrence of a non-crossover chromosome, but the low incidence frequency of such an event has impaired the development of completely non-recombinant lines (*51*).

Even though crossover recombination is considered to be essential for proper pairing and segregation of the homologous chromosomes during meiosis, it is possible to obtain genotypes containing non-crossover chromosomes in standard populations. This is because a single crossover in a homologous chromosome pair is enough to ensure proper segregation of the homologs, and in that case only two of the four chromatids actually recombine (*4*). The other two non-crossover chromatids segregate normally in the resulting gametes after meiosis II.

However, the chance that a single gamete contains only non-crossover chromatids rapidly becomes smaller with increasing numbers of chromosomes. Additionally, since the occurrence of non-crossover chromatids is strongly linked with chromosome size (among other factors), non-crossover chromatids of large chromosomes are less likely to occur than those of smaller chromosomes. Still, completely non-recombinant CSLs can be obtained without transgenic techniques by clever backcross strategies in which one selects for non-crossover chromosomes (*52*).

Chromosome substitution lines as a foundation for genetic mapping

Attempts to obtain CSLs through traditional backcrossing were previously also made in Arabidopsis, although this has never led to the generation of more than a few CSLs (*52*) due to the restrictions mentioned above. However in other species larger CSL panels have been successfully constructed, even though still limited to CSLs with a single chromosome substitution (sCSLs) (*53-56*).

The initial idea of CSLs as a genetic resource should be credited to Sears, who already created an entire panel of monosomic wheat genotypes in 1954 (*57*). In this panel each monosomic line missed one chromosome of either of the three genomes of wheat. In the same publication Sears coined the idea to intercross and continuesly backcross such monosomic plants to create chromosomal substitutions. The first sCSL panels derived from intervarietal crosses using these wheat monosomic genotypes followed shortly thereafter in 1957 (*55*).

These sCSL panels have thereafter been broadly applied to identify chromosomal effects in plant height, root growth and ear-emergence, and have even been instrumental in identifying the causal genes for seed size, vernalization and frost tolerance (*58, 59*). In the late nineties the experimental analyses of CSLs in wheat diminished due to the emerging possibility of constructing new mapping resources with the advent of DNA markers. Also using DNA markers, it became apparent that not all CSLs were correctly identified as such and as a result, due to the low number of backcrosses executed, genetic effects could segregate in further fine-mapping experiments (*60*).

Almost at the same time as in wheat, CSLs were developed in *Drosophila melanogaster* (*56*), and later CSLs were mentioned as possible mapping resource in other species as well (*51, 61*). However, the production of CSLs remained problematic due to the long time required to generate such lines, and the uncertainty of proper non-crossover occurrence, causing the panels to consist of only either a few sCSLs or a set of sCSLs in a single genetic background (unless stated otherwise background refers to the genetic composition of the remainder of the nuclear genotype). With the aid of new genotyping technology, the drawbacks of creating CSLs were overcome and during the beginning of the twenty-first century CSLs gained renewed interest in several model species like mice, rat, and Arabidopsis (*52-54*).

Analysis of genetic effects in chromosome substitution lines

By partitioning the genome of CSLs into clearly delineated building blocks, the effect of genotypic variation at each single chromosome and each combination of chromosomes can be discretely and accurately estimated (*51*). Especially the sCSL panels developed in rodents have been used extensively during the last two decades, identifying high numbers of QTLs (or QTC for chromosomes) for a variety of traits (*62-64*). In current existing sCSL panels the effect is tested in only a single recurrent background similar to typical introgression lines and mutant lines (substitution/ introgression/mutant versus reference).

When multiple substitution or introgressions are present in homozygous genotypes analysis of genetic effects do not need to be limited to main effects. Also interaction effects can be estimated between two loci, with four haplotypes (*AABB; AAbb; aaBB; aabb*; **Fig. 3**). In a comparison with other types of mapping resources the sCSL panels were shown to be especially superior in detecting such genetic interactions, i.e. epistasis (*62*).

In 1909 William Bateson was the first to coin the term epistasis (*65*) for the phenomenon of a departure from the expected Mendelian phenotypic segregation ratio of 9:3:3:1 in the F_2 offspring of a dihybrid cross (*AABB x aabb*; where `*A*` and `*B*` denote the different loci, and the capital vs non-capital letters indicate two different founder alleles). His definition thus mainly concerned multigenic but discrete traits that segregate with independent assortment of the involved genes, leading to the masking of a phenotype for a genotypic class (*66*).

A second definition of epistasis came ten years later from Ronald Fisher in 1918, who initially called it 'epistacy'. This definition, which is used throughout this thesis, is most clearly defined as an interaction between two loci that enhance or decrease trait values more than can be expected from their additive effects (**Fig. 3**). Following his definition, Fisher proposed to analyse epistasis as the departure from a specific linear model that describes the relationship between predictive factors (i.e. genetic loci) (*67*).

In the simplest case of a haploid (or homozygous diploid) genetic model this is symbolised as $y = a_1 + a_2 + i_{12} + e$, where y is the observed phenotype, a is the individual (main or additive) effect of each allele at loci 1 and 2, i is the interaction between a_1 and a_2 (this term is excluded in the additive model) and e contains all the error of the model. Thus genetic variation that cannot be explained by the accumulation of additive effects of single loci must therefore be due to genetic interactions between loci, i.e. epistasis.

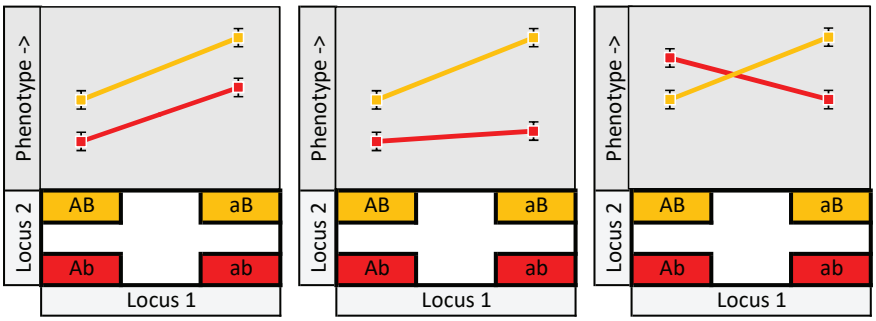


FIGURE 3 | A variety of different genetic models for a two locus model. In all figures the x-axis shows the locus 1 genotypes, and the y-axis the phenotypic value. For each locus position it is possible to have two alleles, represented as capitalized or non-capitalized letters (e.g. at locus 1 `*A*` or `*a*`). In a homozygous diploid species thus only two possible combinations of the two alleles exist; *AA* and *aa*. Considering two loci, four possible genetic combinations can be present. Each panel represents a different two locus model, with the genotypes shown in the bottom diagram. The panels show from left to right an *additive* effect, *quantitative epistasis* (no crossover interaction; positive or negative epistasis) and *qualitative epistasis* (crossover interaction; sign epistasis). In the case of quantitative epistasis it is likely that one or both loci are also detected as a main effect (when the epistasis is of large effect size). In qualitative (or sign) epistasis the effect of an allele in one background is cancelled out in another background. Here it is plausible that both loci remain undetected as significant loci since the average effect of each separate allele usually does not differ.

Complete chromosome substitution panels

During the last couple of years, reports on CSLs in mice with multiple chromosome substitutions have shown the value of locating the genetic factors explaining epistatic interactions on a chromosomal level and have hinted at the opportunities of a complete set of CSLs (*68, 69*).

As of yet no complete panel of all possible single and multiple chromosome substitution lines that would allow to interrogate all possible combinations of pairs of parental chromosomes for epistatic effects has been developed in any species. The lack thereof is mainly attributed to the basic chromosome number of most species. The formula $g = x^n$ can be used to calculate the total number of different homozygous CSL genotypes g, where x represents the number of different founder genotypes for the panel and n the basic haploid chromosome number of the subjected species.

For Arabidopsis, with only five chromosomes, it can be calculated that a complete CSL panel for two founder genotypes, containing all possible combinations of the chromosomes consists of $2^5 = 32$ CSLs. This implies that the number of possible genotypes in such a population is limited, unlike in any other genetic resource. Therefore, one can systematically study all different types of genetic effects with such a panel of CSLs, i.e. additive, and epistatic effects in homozygous genotypes and dominance effects when additional heterozygous genotypes are obtained from intercrossing CSLs (*51, 57, 70*). Such a panel of all different possible CSL genotypes of a biparental cross might thus serve as an ideal starting point for the mapping of complex traits, even though the resolution is restricted to complete chromosomes. In **Chapter 3** the first complete CSL panel of Arabidopsis -or any other species- is presented.

CSL panels containing all possible genotypes can be considered more similar to other typical homozygous segregating populations, such as RILs and DHs, where multiple loci segregate simultaneously in the genetic background. A major drawback of these conventional populations is that large effect QTLs are usually detected using CIM in standard biparental populations, but small effect QTLs remain problematic to identify (*3*). Even more so due to the high recombination frequency, QTLs that interact with other QTLs and which are thus dependent on multiple loci (and their allele frequencies in the sample population), will be even harder to detect and usually end up inflating the error term.

The estimation of chromosome effects in a complete CSL panel with statistical analyses can be performed using a regression model where the parental (founder) origin of each chromosome is included as a single explanatory variable (*51*). In a panel of only homozygous genotypes, all genetic effects can be attributed to either

additive single chromosome effects (QTC_a), or inter-chromosomal effects (QTC_i). In contrast to standard biparental populations, with a complete CSL panel all possible genetic effects can in principle be modelled and estimated simultaneously at the resolution of whole chromosomes, including the interactions (epistasis), ensuring that the error term contains no QTC background variation. The addition of complete CSL panels, including all different chromosome substitutions lines, to the current mapping populations thus have the advantage that all different kind of genetic effects can be estimated with relative ease.

Additional considerations for chromosome substitution lines

The fact that CSLs are homozygous provides equal advantages over heterozygous populations as other homozygous populations do in terms of immortality of the population. However, because of the small population size of a CSL panel in Arabidopsis, genotype replication can be performed to a higher degree than usual, reducing experimental error and subsequently increasing the detection power of chromosome effects (*3*).

Homozygosity of the genotypes allows not only replication in a single experiment, but also replication of an entire experiment or even in different environments. This allows to quantify the influence of the environment on different genotypes (Genotype-by-Environment interactions; GxE), or on specific QTLs (QTLxE) (*71*). Designated statistical software has been developed to detect such interacting QTLs and especially in a plant breeding setting where crop varieties are bred for a specific environment, detection of GxE can be hugely beneficial. Considering the small population size of a complete CSL panel, these can easily be used in multiple experiments with high replicate numbers within experiments.

The downside of CSLs is that effects can only be allocated to the level of an entire chromosome, and thus the detected effect can be caused by several linked QTLs that together determine the observed phenotype (*51*). Although this seems like a major drawback, conventional segregating populations also initially often do not provide confidence intervals smaller than entire chromosome arms and follow-up analyses require large population sizes and dense genotypic data to pinpoint QTL effects to smaller genomic regions (*3*). Therefore, QTL fine-mapping remains an iterative process where frequently a single locus appears to contain multiple genes with additive and/or epistatic effects on the phenotype (*72*).

Moreover, CSLs can also serve as the perfect start for fine-mapping any detected genetic effect. Since all different combinations of chromosomal substitutions are available, the contribution of the QTC in each genetic background can be estimated,

which allows selection of the CSLs that show the largest phenotypic contrast that can serve as the parents for a segregating population. Using this approach, NILs can be quickly generated for fine-scale mapping and validation of the QTLs. CSLs are thus not only a valuable complement to the array of current mapping resources in a high-throughput-genotyping era, but also provide tools for further investigations.

OUTLINE OF THIS THESIS

First, in **Chapter 2** the development of a typical doubled haploid population from a cross between two late flowering (vernalization-requiring) Arabidopsis accessions is described. It is shown that with the use of the *GFP-tailswap* haploid inducer line the production of immortal doubled haploids from representatives of the winter annuals, a group of accessions that has been mainly excluded in most biparental mapping populations due to their long generation time, is shortened to only three generations. This makes it feasible to investigate the unexplored natural diversity for quantitative traits hidden in these accessions. Additionally, since the production of doubled haploids (diploids) with the genome elimination mutant requires the generation of haploids (monoploids), this offers the unique possibility to investigate non-additive ploidy effects. This is the first description of an extensive genetic mapping approach for studying this kind of genetic effects.

Chapter 3 presents the development of the first complete set of chromosome substitution lines (CSLs) for any organism. This genetic resource makes the study of complex genetic traits as comprehensible as possible. In the CSL only 5 loci (i.e. whole chromosomes) segregate, making this the mapping population with the simplest possible genetic architecture. Additionally a set of near-isogenic doubled haploid lines that segregate for only a single chromosome were developed to complement the low resolution of the original CSL panel. Using these resources, it is illustrated that a complete CSL panel can be a remarkably powerful genetic resource to detect not only main effects caused by chromosomal substitutions, but also to detect epistasis in a systematic and unbiased way.

Where in **Chapter 3** the study was limited to the detection of chromosome effects in only two quantitative traits, **Chapter 4** describes how such a CSL panel can be used in an -omics approach. This is the first genetical-proteomics approach where hundreds of proteins are detected and quantified. Even when reducing the population size of the complete panel to about half its size additive effects and two-way interactions in a specific genetic background can be detected for individual proteins. With the plethora of protein intensities this approach also allows a more generalized estimate of the contribution of epistasis to genetic variance.

Chapter 5 describes a new method to quickly obtain CSLs from wild-type Arabidopsis hybrids by exploring a new way to execute reverse breeding, which previously used a RNAi transgene to suppress meiotic crossover recombination. First, it is shown how crossover recombination can be strongly, and transiently reduced by the use of virus-induced gene silencing (VIGS). This means that gametes without CO recombination can be much easier obtained from F_1 hybrids. Second, the transient down-regulation of *MUT S HOMOLOGUE 5* (*MSH5*) instead of *DMC1* leads to the production of low-crossover spores that can be regenerated as doubled haploid offspring. This not only increases the proportion of viable gametes produced by the hybrid, but also allows the technique to be applied in species with higher chromosome numbers.

Finally, **Chapter 6** discusses the implications of the findings in the previous chapters for the scientific community at large and what future endeavours can be undertaken with the populations presented here. Additionally, compelling evidence is presented to the breeding community that this approach is a valuable asset for future genetic mapping purposes.

REFERENCES

1. J. Bergelson, F. Roux, Towards identifying genes underlying ecologically relevant traits in Arabidopsis thaliana. *Nature Reviews Genetics* 11, 867-879 (2010).
2. T. Mitchell-Olds, J. Schmitt, Genetic mechanisms and evolutionary significance of natural variation in Arabidopsis. *Nature* 441, 947-952 (2006).
3. M. Lynch, B. Walsh, *Genetics and analysis of quantitative traits*. (Sinauer Sunderland, MA, 1998), vol. 1.
4. R. Mercier, C. Mezard, E. Jenczewski, N. Macaisne, M. Grelon, The molecular biology of meiosis in plants. *Annu Rev Plant Biol* 66, 297-327 (2015).
5. P. A. Salome *et al.*, The recombination landscape in Arabidopsis thaliana F2 populations. *Heredity* 108, 447-455 (2012).
6. S. D. Tanksley, Mapping Polygenes. *Annual Review of Genetics* 27, 205-233 (1993).
7. J. N. Fitz Gerald *et al.*, New Arabidopsis Advanced Intercross Recombinant Inbred Lines Reveal Female Control of Nonrandom Mating. *Plant Physiology* 165, 175-185 (2014).
8. M. D. McMullen *et al.*, Genetic Properties of the Maize Nested Association Mapping Population. *Science* 325, 737-740 (2009).
9. J. Cockram, I. Mackay, in *Plant Genetics and Molecular Biology,* R. K. Varshney, M. K. Pandey, A. Chitikineni, Eds. (Springer International Publishing, Cham, 2018), pp. 109-138.
10. M. Koornneef *et al.*, Linkage map of Arabidopsis thaliana. *Journal of Heredity* 74, 265-272 (1983).
11. S. E. Levy, R. M. Myers, Advancements in Next-Generation Sequencing. *Annual review of genomics and human genetics* 17, 95-115 (2016).
12. E. S. Lander, D. Botstein, Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* 121, 185-199 (1989).
13. C. Shindo, G. Bernasconi, C. S. Hardtke, Natural genetic variation in arabidopsis: Tools, traits and prospects for evolutionary ecology. *Annals of Botany* 99, 1043-1054 (2007).
14. C. Bazakos, M. Hanemian, C. Trontin, J. M. Jiménez-Gómez, O. Loudet, New strategies and tools in quantitative genetics: How to go from the phenotype to the genotype. *Annual Review of Plant Biology* 68, 435-455 (2017).
15. S. Atwell *et al.*, Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature* 465, 627-631 (2010).
16. P. X. Kover *et al.*, A multiparent advanced generation inter-cross to fine-map quantitative traits in Arabidopsis thaliana. *PLOS Genetics* 5, (2009).
17. E. A. Greene *et al.*, Spectrum of Chemically Induced Mutations From a Large-Scale Reverse-Genetic Screen in Arabidopsis. *Genetics* 164, 731-740 (2003).
18. R. Kooke, E. Wijnker, J. J. B. Keurentjes, in *Methods in Molecular Biology*. (2012), vol. 871, pp. 3-16.
19. R. S. Fletcher *et al.*, Development of a next-generation NIL library in Arabidopsis thaliana for dissecting complex traits. *BMC Genomics* 14, (2013).
20. J. J. B. Keurentjes *et al.*, Development of a near-isogenic line population of Arabidopsis thaliana and comparison of mapping power with a recombinant inbred line population. *Genetics* 175, 891-905 (2007).
21. C. Lister, C. Dean, Recombinant inbred lines for mapping RFLP and phenotypic markers in Arabidopsis thaliana. *Plant J* 4, 745-750 (1993).
22. M. W. Horton *et al.*, Genome-wide patterns of genetic variation in worldwide Arabidopsis thaliana accessions from the RegMap panel. *Nature Genetics* 44, 212-216 (2012).
23. P. A. Salomé *et al.*, The recombination landscape in Arabidopsis thaliana F 2 populations. *Heredity* 108, 447-455 (2012).
24. P. A. Salomé *et al.*, Genetic architecture of flowering-time variation in Arabidopsis thaliana. *Genetics* 188, 421-433 (2011).
25. B. J. Vilhjálmsson, M. Nordborg, The nature of confounding in genome-wide association studies. *Nature Reviews Genetics* 14, 1-2 (2013).
26. B. Brachi *et al.*, Linkage and association mapping of Arabidopsis thaliana flowering time in nature. *PLOS Genetics* 6, 40 (2010).
27. S. Balasubramanian *et al.*, QTL mapping in new Arabidopsis thaliana advanced intercross-recombinant inbred lines. *PLoS ONE* 4, (2009).

28. M. V. Rockman, L. Kruglyak, Breeding designs for recombinant inbred advanced intercross lines. *Genetics* 179, 1069-1078 (2008).
29. X. Huang *et al.*, Analysis of natural allelic variation in Arabidopsis using a multiparent recombinant inbred line population. *Proceedings of the National Academy of Sciences of the United States of America* 108, 4488-4493 (2011).
30. J. M. Dunwell, Haploids in flowering plants: Origins and exploitation. *Plant Biotechnology Journal* 8, 377-424 (2010).
31. M. Wędzony *et al.*, in *Advances in Haploid Production in Higher Plants,* A. Touraev, B. P. Forster, S. M. Jain, Eds. (Springer Netherlands, Dordrecht, 2009), pp. 1-33.
32. M. Koornneef, D. Meinke, The development of Arabidopsis as a model plant. *Plant Journal* 61, 909-921 (2010).
33. D. Weigel, Natural variation in arabidopsis: From molecular genetics to ecological genomics. *Plant Physiology* 158, 2-22 (2012).
34. M. Ravi, S. W. L. Chan, Haploid plants produced by centromere-mediated genome elimination. *Nature* 464, 615-618 (2010).
35. J. Gil-Humanes, F. Barro, in *Advances in Haploid Production in Higher Plants,* A. Touraev, B. P. Forster, S. M. Jain, Eds. (Springer Netherlands, Dordrecht, 2009), pp. 65-73.
36. K. J. Kasha, K. N. Kao, High Frequency Haploid Production in Barley (Hordeum vulgare L.). *Nature* 225, 874 (1970).
37. R. E. Clausen, M. C. Mann, Inheritance in Nicotiana Tabacum. *V. The Occurrence of Haploid Plants in Interspecific Progenies* 10, 121-124 (1924).
38. H. G. Tunner, S. Heppich, Premeiotic genome exclusion during oogenesis in the common edible frog,Rana esculenta. *Naturwissenschaften* 68, 207-208 (1981).
39. K. J. Kasha, K. N. Kao, High Frequency Haploid Production in Barley (Hordeum vulgare L.). *Nature* 225, 874-876 (1970).
40. M. Sanei, R. Pickering, K. Kumke, S. Nasuda, A. Houben, Loss of centromeric histone H3 (CENH3) from centromeres precedes uniparental chromosome elimination in interspecific barley hybrids. *Proceedings of the National Academy of Sciences* 108, E498-E505 (2011).
41. M. Ravi *et al.*, The Rapidly Evolving Centromere-Specific Histone Has Stringent Functional Requirements in Arabidopsis thaliana. *Genetics* 186, 461-471 (2010).
42. L. Comai, Genome Elimination: Translating Basic Research into a Future Tool for Plant Breeding. *PLOS Biology* 12, e1001876 (2014).
43. C. L. Wijnen, J. J. B. Keurentjes, Genetic resources for quantitative trait analysis: novelty and efficiency in design from an Arabidopsis perspective. *Current Opinion in Plant Biology* 18, 103-109 (2014).
44. M. P. A. Marimuthu *et al.*, Synthetic Clonal Reproduction Through Seeds. *Science* 331, 876-876 (2011).
45. M. Ravi *et al.*, A haploid genetics toolbox for Arabidopsis thaliana. *Nature Communications* 5, 5334 (2014).
46. E. Wijnker *et al.*, Reverse breeding in Arabidopsis thaliana generates homozygous parental lines from a heterozygous plant. *Nature Genetics* 44, 467-470 (2012).
47. D. K. Seymour *et al.*, Rapid creation of Arabidopsis doubled haploid lines for quantitative trait locus mapping. *Proceedings of the National Academy of Sciences* 109, 4227-4232 (2012).
48. N. Fulcher, K. Riha, Using Centromere Mediated Genome Elimination to Elucidate the Functional Redundancy of Candidate Telomere Binding Proteins in Arabidopsis thaliana. *Frontiers in Genetics* 6, (2016).
49. R. Dirks *et al.*, Reverse breeding: a novel breeding approach based on engineered meiosis. *Plant Biotechnol J* 7, 837-845 (2009).
50. E. Wijnker *et al.*, Hybrid recreation by reverse breeding in Arabidopsis thaliana. *Nature Protocols* 9, 761-772 (2014).
51. J. H. Nadeau, J. B. Singer, A. Matin, E. S. Lander, Analysing complex genetic traits with chromosome substitution strains. *Nature Genetics* 24, 221-225 (2000).
52. R. Koumproglou *et al.*, STAIRS: a new genetic resource for functional genomic studies of Arabidopsis. *Plant J* 31, 355-364 (2002).
53. J. B. Singer *et al.*, Genetic dissection of complex traits with chromosome substitution strains of mice. *Science* 304, 445-448 (2004).

54. A. W. Cowley, M. Liang, R. J. Roman, A. S. Greene, H. J. Jacob, Consomic rat model systems for physiological genomics. *Acta Physiologica Scandinavica* 181, 585-592 (2004).
55. J. Kuspira, J. Unrau, Genetic analyses of certain characters in common wheat using whole chromosome substitution lines. *Canadian Journal of Plant Science* 37, 300-326 (1957).
56. M. B. Seiger, The effects of chromosome substitution on male body weight of *Drosophila melanogaster*. *Genetics* 53, 237-248 (1966).
57. E. R. Sears, The aneuploids of common wheat. *Research Bulletin 572*, (1954).
58. P. I. Payne, C. N. Law, E. E. Mudd, Control by homoeologous group 1 chromosomes of the high-molecular-weight subunits of glutenin, a major protein of wheat endosperm. *Theoretical and Applied Genetics* 58, 113-120 (1980).
59. G. Galiba, S. A. Quarrie, J. Sutka, A. Morgounov, J. W. Snape, RFLP mapping of the vernalization (Vrn1) and frost resistance (Fr1) genes on chromosome 5A of wheat. *Theoretical and Applied Genetics* 90, 1174-1179 (1995).
60. T. Ryu Endo, B. S. Gill, Somatic karyotype, heterochromatin distribution, and nature of chromosome differentiation in common wheat, Triticum aestivum L. em Thell. *Chromosoma* 89, 361-369 (1984).
61. S. S. Banga, C-genome chromosome substitution lines in Brassica juncea (L.) Coss. *Genetica* 77, 81-84 (1988).
62. D. A. Buchner, J. H. Nadeau, Contrasting genetic architectures in different mouse reference populations used for studying complex traits. *Genome Research* 25, 775-791 (2015).
63. H. Shao *et al.*, Genetic architecture of complex traits: Large phenotypic effects and pervasive epistasis. *Proceedings of the National Academy of Sciences* 105, 19910-19914 (2008).
64. S. H. Spiezio, T. Takada, T. Shiroishi, J. H. Nadeau, Genetic divergence and the genetic architecture of complex traits in chromosome substitution strains of mice. *BMC Genetics* 13, 38 (2012).
65. W. Bateson, *Mendel's Principles of Heredity*. (Cambridge University, 1909).
66. P. C. Phillips, The Language of Gene Interaction. *Genetics* 149, 1167-1171 (1998).
67. H. J. Cordell, Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics* 11, 2463-2468 (2002).
68. J. P. Rapp, M. R. Garrett, A. Y. Deng, Construction of a double congenic strain to prove an epistatic interaction on blood pressure between rat chromosomes 2 and 10. *The Journal of clinical investigation* 101, 1591-1595 (1998).
69. A. Chen, Y. Liu, S. M. Williams, N. Morris, D. A. Buchner, Widespread epistasis regulates glucose homeostasis and gene expression. *PLOS Genetics* 13, e1007025 (2017).
70. A. E. Hill, E. S. Lander, J. H. Nadeau, in *Cardiovascular Disease: Methods and Protocols Volume 1: Genetics,* Q. K. Wang, Ed. (Humana Press, Totowa, NJ, 2007), pp. 153-172.
71. M. P. Boer *et al.*, A Mixed-Model Quantitative Trait Loci (QTL) Analysis for Multiple-Environment Trial Data Using Environmental Covariables for QTL-by-Environment Interactions, With an Example in Maize. *Genetics* 177, 1801-1813 (2007).
72. J. Flint, W. Valdar, S. Shifman, R. Mott, Strategies for mapping and cloning quantitative trait genes in rodents. *Nature Reviews Genetics* 6, 271-286 (2005).

Chapter 2

Detection of genotype-by-ploidy effects in a mono- and diploid mapping population of *Arabidopsis thaliana*

Cris L. Wijnen[1,2], Frank F.M. Becker[1], Andries A. Okkersen[1], C. Bastiaan de Snoo[3], Martin P. Boer[2], Fred A. van Eeuwijk[2], Erik Wijnker[1], Joost J.B. Keurentjes[1]

[1] Laboratory of Genetics, Wageningen University & Research, Droevendaalsesteeg 1, 6708 PB, Wageningen, The Netherlands.
[2] Biometris, Wageningen University & Research, Droevendaalsesteeg 1, 6708 PB, Wageningen, The Netherlands.
[3] Rijk Zwaan R&D Fijnaart, Eerste kruisweg 9, 4793 RS Fijnaart, The Netherlands.

ABSTRACT

It has previously been shown that plant accessions can show different phenotypic responses following polyploidization, but such ploidy-dependent phenotypic differences were never mapped to genomic regions. To map such effects, one requires segregating populations at different ploidy levels. The availability of an efficient haploid-inducer line in Arabidopsis allows for the rapid development of large segregating haploid offspring populations. Because Arabidopsis haploids can be self-fertilised to give rise to homozygous doubled haploids, the same plants can be phenotyped at both haploid and diploid ploidy levels. We compared haploid and diploid phenotypes of offspring derived from a cross between two late flowering accessions to map gene x ploidy (GxP) interactions. We demonstrate the detection of ploidy specific QTLs at both ploidy levels. This implies that more QTLs will be detected when phenotypic measurements of the monoploids are included in a combined QTL analyses. A multi-trait analysis indicated pleiotropic effects for a number of the ploidy specific QTLs, including opposite effects at different ploidy levels in multiple traits. Taken together, we provide evidence that genetic variation between different Arabidopsis accessions is causal for differences in phenotypic responses to altered ploidy levels, revealing a GxP effect. Additionally, by comparing doubled haploids with and without vernalization, we revealed a major vernalization specific quantitative trait locus (QTL) for variation in flowering time, which is strongly linked to the vernalization insensitive gene *VIN3*. *VIN3* mediates the initial transcriptional repression of the homeotic gene *FLC*, a floral repressor, after cold treatment.

INTRODUCTION

Phenotypic effects caused by different ploidies have so far been very elusive and difficult to study in *Arabidopsis thaliana*. Nonetheless, the impact of ploidy is illustrated by the strong effect on quantitative traits such as salt and drought tolerance, and relative growth rate (*1-3*). Most attempts to reveal ploidy effects have used naturally occurring autotetraploids such as Warschau (Wa-1) (*4, 5*) or artificially induced tetraploid accessions (*6*), which were compared to their diploid and triploid counterparts (*3, 4*). While a biparental mapping population was developed using Wa-1 as one of the parental genotypes, it was only later discovered that this genotype was tetraploid and that the inbred lines were derived from triploids (*4, 7*). Therefore, in this population the ploidy level segregated and many of the genotypes are not explicitly diploid or tetraploid according to flowcytometry (*1, 4*). Notwithstanding this unstable population, a mapping resource of different stable different ploidy levels has not been specifically developed in *Arabidopsis*.

Monoploids (individuals consisting of somatic cells containing only the basic number of chromosomes) are usually not taken into account in studies that investigate the effect of ploidy, although exceptions are studies in maize (*8*), yeast (*9*), potato (*10*) and Chinese cabbage (*11*) in which ploidy series including monoploids were compared. These studies focus on transcriptional changes induced by ploidy in only a single or a few genotypes. For instance, Stupar *et al.*Stupar, Bhaskar, Yandell, Rensink, Hart, Ouyang, Veilleux, Busse, Erhardt, Buell and Jiang (*10*) demonstrated that more than 50% of the analysed transcripts displayed expression differences between monoploids and diploids or tetraploids, suggesting large developmental differences between plants of different ploidy levels.

The discovery of a haploid inducer line in Arabidopsis now allows the quick generation of haploid lines from diploid individuals (*12*). The generation of haploids occurs through elimination of the haploid inducer genome in the offspring of a cross between a haploid inducer with a wild type diploid. Diploid Arabidopsis somatic cells contain 2n = 2x = 10 chromosomes. Here, haploids that are generated with the haploid-inducer contain n = x = 5 chromosomes and thus are equivalent to monoploids. These monoploids are mostly sterile and cannot be maintained as such. The haploid plants do however set seed and give rise to homozygous doubled haploids (DHs) because of either somatic doubling of cell lines in the haploid plants or the incidental non-disjunction of all homologs at meiosis I. While the maternally derived seed coat contains the maternal ploidy level, the embryo and endosperm (2*n* = 2*x* and *n* = 3*x,* respectively) contain equal chromosome numbers in seeds derived from mono- or diploids. Doubled haploids thus contain a doubled genome, and consist again of diploid somatic cells containing 2n = 2x = 10 chromosomes.

When the haploid inducer is crossed with an F_1 hybrid of two distinct accessions, the recombinant gametes of the hybrid are the sole source of the resulting monoploid offspring genomes. By allowing the monoploids to produce DH seeds, the monoploid genome is immortalized in homozygous diploids. The production of such a diploid mapping population using genome elimination thus has the advantage that initially large amounts of segregating monoploids are produced, which except for the ploidy level are genetically identical to the DH obtained in the next generation (*13, 14*). These monoploids may provide a useful resource for genetic mapping and allow assessment of ploidy effects in comparisons with their subsequent isogenic diploid offspring.

The generation of DH mapping populations has an advantage over the more commonly used Recombinant Inbred Lines (RILs). RILs are generated though repeated self-fertilization of an F1 individual by single seed descent, to give rise to homozygous populations after eight-to-ten generations of inbreeding. This contrasts to DHs for which homozygous diploid populations from an F1 can be obtained in only three generations (*13, 15, 16*). This fast development advantage of DH populations allows the investigation of natural variation in late-flowering winter annual accessions, whereas for most existing experimental biparental mapping populations either one or two summer annual parental accessions are used, which shorten the generation time due to their early-flowering phenotype (*16-20*).

Summer annuals germinate in spring and flower within a short period of time, while winter annuals germinate in autumn, survive winter as a rosette and usually flowering is induced by vernalization, a period of cold conditions. Without a cold period flowering time of winter annuals usually takes at least a few month, but can be severely shortened in an experimental setting by vernalization. Until a decade ago, homozygous mapping resources often involved common laboratory strains such as Columbia (Col-0), Landsberg *erecta* (L*er*) or Cape Verde Islands (CVI), all summer annual accessions (*16-18*). The last ten years several additional resources have become available, including more accessions to represent the global diversity of the species, even though a bias towards the use of early-flowering plants remains (*18-20*).

Although the haploid-inducer approach eliminates the need for a lengthy inbreeding process to obtain homozygous lines, the only three DH populations reported for Arabidopsis also originate from early-flowering accessions (*13, 21, 22*). However, the short time needed to produce DH populations through genome elimination is ideally suited for winter annual accessions with late-flowering phenotypes. The complementation of the many early-flowering mapping populations with novel late-flowering ones would do justice to the high genetic diversity present in the species Arabidopsis.

Here, we describe the development and phenotyping of a mono- and subsequent diploid population from a cross between the two late-flowering accessions, T540 (Kävlinge, Sweden) and Ge-0 (Geneva, Switzerland). These accessions display large phenotypic differences in biomass formation and flowering time after vernalization. These quantitative traits were investigated for three different aspects, first we tested for the presence of a genotype-by-environment effect by comparing flowering time of the diploid generation with and without vernalization. We demonstrate that exploiting genetic variation in late-flowering accessions can increase our knowledge even in a well-studied trait like flowering time. Secondly, we investigated the possibility of detecting genotype-by-ploidy (GxP) interactions by performing a combined analysis across monoploids and diploids, using a multi-trait QTL model approach. Here we were able to detect QTLs for the selected traits in only one of the two ploidy levels, and reveal genotype-by-ploidy interactions based on the QTL effect size. This demonstrates that including phenotypic data of a monoploid generation during a QTL analyses, may have an added benefit for detecting QTLs. Finally, we analysed all traits together to detect pleiotropic QTLs, and show that most detected QTLs influence multiple phenotypes in both generations (i.e. ploidy levels), while a minor number of QTLs affect predominantly a single trait at a specific ploidy level. Taken together, this study advocates the use of both the mono- and diploid generations during haploid production to detect additional QTLs that might have remained undetected otherwise.

RESULTS

Development and phenotyping of a mono- and its derived diploid mapping population

To explore the effect of ploidy on genetic mapping in Arabidopsis, a segregating population was generated from a cross between two late flowering accessions, T540 and Ge-0 (**Fig. 1**). Briefly, the late flowering accessions T540 and Ge-0 were crossed to produce an F_1 hybrid. This hybrid was subsequently manually crossed to a haploid inducer line *GFP-tailswap* (12), from which approximately 250 seeds were obtained. These seeds were stratified and pre-germinated, after which seedlings were transferred to Rockwool and grown for three weeks under long day conditions in a climate controlled growth chamber. After visual inspection, 210 potentially haploid plants were transferred to a cold room for eight weeks vernalization under short-day conditions. Once vernalized, plants were transferred to a greenhouse under long day conditions and subsequently formed inflorescences, flowered and set seeds. At the end of the growth period non-destructive phenotypes were measured, i.e. main

stem length, branching from rosette and branching from the main inflorescence, allowing the monoploids to produce doubled haploid seeds. These seeds formed the subsequent diploid generation. The diploid seeds harvested from monoploid plants were also analysed for average seed size.

In a second experiment the 210 potential DH lines were grown in a climate chamber under similar conditions as described for the monoploids. Ten replicates of each line of the diploid population were grown in a completely randomised design. After three weeks, five of these were transferred to a greenhouse to record the time to flowering. The other five replicates were also transferred after three weeks, but to a cold room and vernalized for eight weeks at 4 °C. These plants were thereafter transferred back to the climate chamber with long day conditions and phenotyped for flowering time and for the same traits as the monoploids were phenotyped for. Assuming all replicate plants were isogenic, one plant of each genotype was selected for genotyping, which was successful for 195 lines. After data analysis of genotypes and phenotypes, 171 genotypes, for which phenotypic data at both ploidy levels could be obtained, were selected. The phenotypic data of these lines were used for all further analyses. The genotype data of these lines were used for the construction of a genetic map and the QTL mapping of the analysed traits using standard methods.

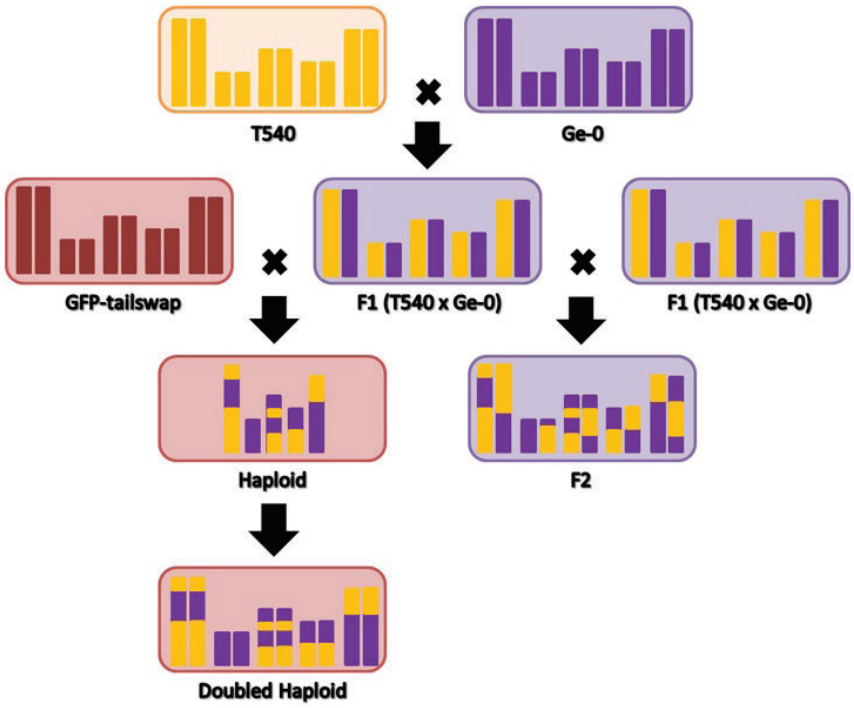


FIGURE 1 | The crossing scheme for development of the different populations. Each parental genotype (T540 yellow; Ge-0 purple) is depicted by five double vertical bars, which represent the five chromosomes, while the box indicates the respective genotype of the cytoplasm. The haploid is a line of a Col-0 genotypic background (red). Note that the haploids (monoploids) and doubled haploids (diploids) retain the cytoplasm of the haploid inducer line, while the F2 population retains the cytoplasm of the original F1.

In addition to the artificial haploid and DH mapping populations, a classical F_2 population of 71 lines derived from the same T540 x Ge-0 F_1 hybrid was generated and grown simultaneously with the doubled haploids in the second experiment. These F_2 plants were all subjected to vernalization and phenotyped for flowering time. Moreover, this small F_2 set was genotyped with 123 markers and their linkage patterns were compared with those of the DH population to confirm no anomalies occurred during the DH development (**Material and methods & Fig. S1 & S2**). With the exception of a slight genotype distortion at the top of chromosome 1 in the DH population, no systematic differences were observed between the F_2 and the DH population. Importantly, the genetic maps generated from the two populations displayed an almost identical marker order, consistent with the known physical position of markers.

Detection of vernalization specific flowering time QTLs

While a vernalization treatment can have a large overall phenotypic effect on the morphology and inflorescence structures of late flowering accessions (*23, 24*), mapping experiments in Arabidopsis have focussed on detecting QTLs in either early flowering populations or on mapping specific QTLs involved in vernalization requirement with populations derived from parental accessions differing in this aspect (*17-20, 25*). Here we have the opportunity to compare and map natural variation in flowering time with and without vernalization in a late flowering segregating DH and F_2 population.

Although both parents of the DH population are late flowering they do not per se require vernalization to flower. In our experiment without vernalization T540 flowered on average after 101.8 days after sowing (DAS), while this was 86.5 days for Ge-0 (**Table 1**). With vernalization these accessions flowered on average after 19.1 and 14.2 days after transfer (DAT) from the cold, respectively. For the small F_2 population, which was only measured under vernalized conditions, a similar population average phenotype was obtained as for the DH population (**Table 1**). The variation in flowering time between the two accessions segregated in the diploid populations with a minor transgression in both conditions. Without vernalization, the earliest line of the DH population flowered after 63 DAS, while the latest flowered at 123 DAS. With vernalization the difference between extreme lines reduced to only ten days (12 and 22 DAT, respectively). The correlation in flowering time between the vernalized and non-vernalized plants was positive but far from absolute (R^2 = 0.39).

TABLE 1 | Summary statistics for flowering time in the different populations. Note that the doubled haploid (DH) population is referred to as diploid. F_1 and F_2 flowering times were not determined in non-vernalization conditions. FTv, flowering time after vernalization (days after transfer); FTnv, flowering time without vernalization (days after sowing); s.d., standard deviation; Cv, coefficient of variation.

Trait	Genotype	Mean (n)	s.d.	Min	Max	Cv (%)
FTv	Ge-0	14.2 (29)	1.17	13	17	8.2
	T540	19.1 (22)	1.50	15	22	7.8
	F1	17.1 (18)	0.80	16	18	4.7
	F2	15.3 (63)	1.29	13	18	8.4
	Diploid (DH)	15.9 (171)	1.57	12	22	9.8
FTnv	Ge-0	86.5 (8)	16.27	64	107	18.8
	T540	101.8 (4)	14.08	89	121	13.8
	Diploid (DH)	88.4 (163)	12.09	63	123	13.7

The data for flowering time of the doubled haploids under different vernalization conditions allowed a multi-environment composite interval mapping (CIM) where the effect of vernalization was investigated. Additionally, the F_2 population was screened for QTLs in vernalized conditions in a separate analysis. A total of seven QTLs spread over the genome were detected for variation segregating in the DH population (**Table 2**). Of these seven QTLs, three revealed an interaction with the environment, providing evidence for GxE effects of vernalization. One QTL with a GxE effect was located on chromosome 4 and the other two were detected on chromosome 5. The QTL for flowering time after vernalization in the middle of chromosome 4 had a normalized effect size of 0.33 for the T540 allele, while this QTL was not significant (p-value = 0.653) for the non-vernalized plants. In contrast, both GxE QTLs on chromosome 5 were significant in both environments but with different effect sizes (**Table 2**).

The major QTL at the bottom of chromosome 5 had an additive normalized effect size of 0.65 when the plants were vernalized while this was only 0.19 in the non-vernalized set. This major QTL was also detected in the vernalized F_2 population. The other GxE QTL detected on the middle of chromosome 5 in the DH population had a normalized effect size of 0.43 in non-vernalized plants, while this was significantly lower (0.21) in vernalized plants. Additional QTLs without GxE effects were detected on the middle and bottom of chromosome 1, bottom of chromosome 2 and top of chromosome 3 (**Table 2**).

TABLE 2 | Overview of the QTLs identified with the multi-environment QTL analyses. Positions are shown with support intervals in between brackets. The column "*-log10(p)*" indicates the significance of the QTL for the combined traits, while in the column "*p-value*" the specific p-value for each trait is given. FT_nv, flowering time without vernalization (days after sowing); FT_v, flowering time after vernalization (days after transfer). "*Effect size*" is given as the normalized additive effect of the QTL, where positive values indicate a positive effect of the T540 allele and negative values indicate a Ge-0 high value allele; "*s.e.*" is the standard error of the mean effect; "*%EV*" is the explained variance according to a mixed model. For the F_2 population dominance effects could be calculated, which are indicated in the column "*Type*". Significant effects for each QTL and trait are indicated by bold.

Population	Chromosome	Position (cM)	-LOG10(P)	Trait	P-value	Effect size	s.e.	%EV	Type
Diploid (DH)	I	70.8 (5.2-150.6)	4.5	**FT_nv**	**<0.001**	**0.31**	**0.08**	**9.8**	**-**
				FT_v	**0.005**	**0.16**	**0.06**	**2.4**	**-**
	I	149.5 (5.2-150.6)	4.4	FT_nv	0.391	0.06	0.07	0.3	-
				FT_v	**<0.001**	**0.21**	**0.05**	**4.4**	**-**
	II	80.4 (0.9-93.5)	2.3	FT_nv	0.120	-0.11	0.07	1.1	-
				FT_v	**0.002**	**-0.15**	**0.05**	**2.3**	**-**
	III	5.9 (2.9-117.2)	4.9	**FT_nv**	**<0.001**	**0.26**	**0.07**	**6.8**	**-**
				FT_v	**<0.001**	**0.16**	**0.05**	**2.7**	**-**
	IV	61 (5.3-92.8)	10.9	FT_nv	0.653	-0.03	0.07	0.1	-
				FT_v	**<0.001**	**0.33**	**0.05**	**10.9**	**-**
	V	61.2 (2.4-133.1)	9.7	**FT_nv**	**<0.001**	**-0.43**	**0.07**	**18.7**	**-**
				FT_v	**<0.001**	**-0.21**	**0.05**	**4.3**	**-**
	V	121.7 (75.4-133.1)	38.7	**FT_nv**	**0.004**	**0.19**	**0.07**	**3.7**	**-**
		0 (0-0)		**FT_v**	**<0.001**	**0.65**	**0.05**	**42.6**	**-**
F2	IV	5.3 (5.3-92.8)	3.2	**FT_v**	**-**	**0.26**	**0.19**	**2.0**	**Additive**
					-	**0.98**	**0.27**	-	**Dominance**
	V	126.4 (109.5-133.1)	5.7	**FT_v**	**-**	**1.21**	**0.23**	**44.0**	**Additive**
					-	-	-	-	Dominance

Detection of ploidy-related phenotypic differences

To investigate if differences in ploidy also affect the phenotypes of individuals of a population, the measurements of various morphological traits in the mono- and diploid generation were compared (**Fig. 2 and Table 3**). Monoploids were on average much taller than their diploid counterparts (65 versus 46 cm, respectively; **Fig. 2A, Fig. S3 & Table 3**). Illustrative for this difference in length is that more than 60% of the monoploids grew taller than the tallest diploid, which measured 61 cm.

When comparing the monoploids to the diploids a larger average number of branches sprouting from the rosette was observed in the monoploids and like for main stem length, the monoploids displayed more variation,(**Fig. 2B; Table 3**). In addition, branching from the rosette occurred much more frequently in monoploids (95.8%) than in diploids (34.9%). Moreover, a maximum of only three rosette branches was observed in diploids, while monoploids developed on average seven branches from the rosette, with an exceptional maximum of twenty-three branches. In contrast to variation in main stem length and branching from

TABLE 3 | Summary statistics for phenotypes measured in both the mono- and diploid generation. Note that the doubled haploid (DH) population is referred to as diploid. MSL, main stem length (cm); BFR, branching from rosette (nr.); BFS, branching from stem (nr.); SA, seed area (mm2); s.d., standard deviation; Cv, coefficient of variation.

Trait	Population	Mean (n)	s.d.	Min	Max	Cv (%)
MSL	Ge-0	46.6 (22)	5.4	35.5	55.5	11.5
	T540	44.8 (16)	4.8	36.5	53	10.8
	F1	52.9 (17)	3.9	45	62	7.4
	Monoploid	65.1 (168)	9.5	36	97	14.6
	Diploid (DH)	46.5 (169)	4.8	32	61	10.4
BFR	Ge-0	0.7 (27)	1.3	0	4	177.9
	T540	1.7 (20)	1.9	0	6	116.8
	F1	0.7 (13)	1.4	0	4	199.0
	Monoploid	7.4 (165)	4.6	0	23	62.0
	Diploid (DH)	0.4 (169)	0.6	0	3	150.5
BFS	Ge-0	10.3 (28)	1.4	7	13	13.5
	T540	9.1 (20)	1.3	7	11	14.7
	F1	11.2 (18)	0.7	10	12	6.5
	Monoploid	10.1 (165)	2.4	1	16	23.7
	Diploid (DH)	10.1 (171)	1.1	7	12	10.7
SA	Monoploid	0.088 (157)	0.008	0.070	0.107	8.5
	Diploid (DH)	0.085 (164)	0.007	0.063	0.110	8.3

rosette, the variation in branching from the stem spread around almost identical mean values at both ploidy levels, although a larger transgression was observed in the monoploids as compared to the diploids (**Fig. 2C**). Despite the differences in phenotypic variation between the number of branches from the rosette and from the stem, a moderate positive correlation (R^2 = 0.32; **Fig. 3**) could be detected in the monoploids. This resulted in monoploids with up to a total number of thirty-two branches, giving rise to a bushy phenotype (**Fig. S3C**).

Similar to branching from the stem, the phenotypic variation in the size of seeds harvested from mono- or diploid plants centred around a comparable mean for both populations, although the between-line variation was somewhat larger for seeds derived from diploids than those derived from monoploids (**Table 3**). Pearson correlations between mono- and diploids were positive for all traits, but did not reach high values ($0.3 < R^2 < 0.4$; **Fig. 3**), while moderate to high broad sense heritabilities were obtained for most traits segregating in the diploid population ($0.30 < H^2 < 0.83$; **Fig. 3**). This suggests that differences between mono- and diploids can be partly explained by simple additive ploidy effects but that the majority of variation might be the result of more complex genotype-by-ploidy interactions or sampling error.

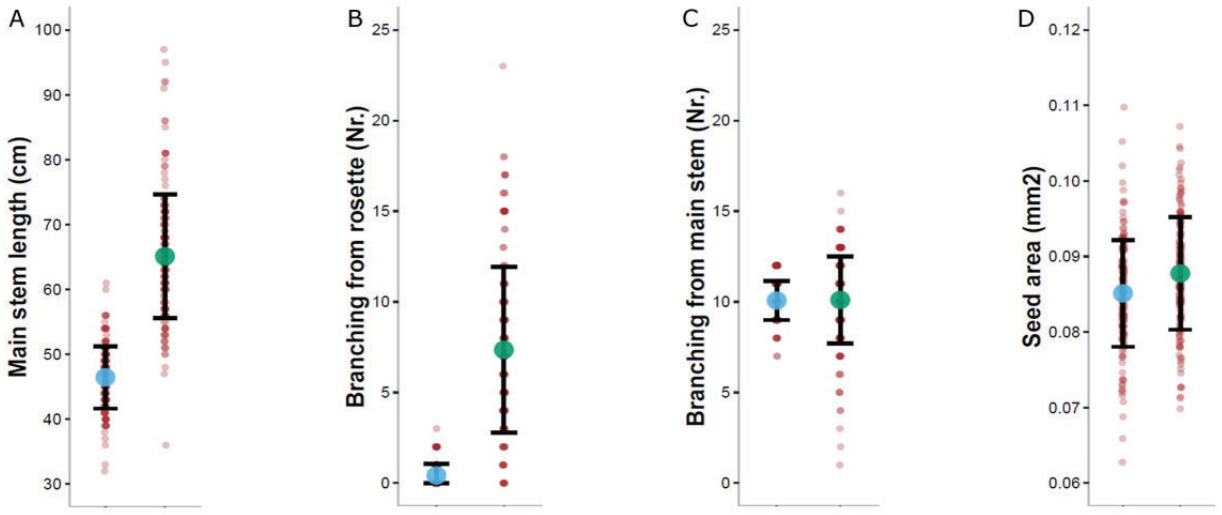


FIGURE 2 | Phenotypic distributions of morphological traits for mono- and diploids. The mean value of the diploids (DH) and monoploids is indicated with the blue and green dots, respectively. The red dots depict the value of the individual monoploids and the line average of five replicates for each diploid, respectively. For seed area only a single measurement was used for the diploids as well. Error bars indicate the standard deviation of the mean.

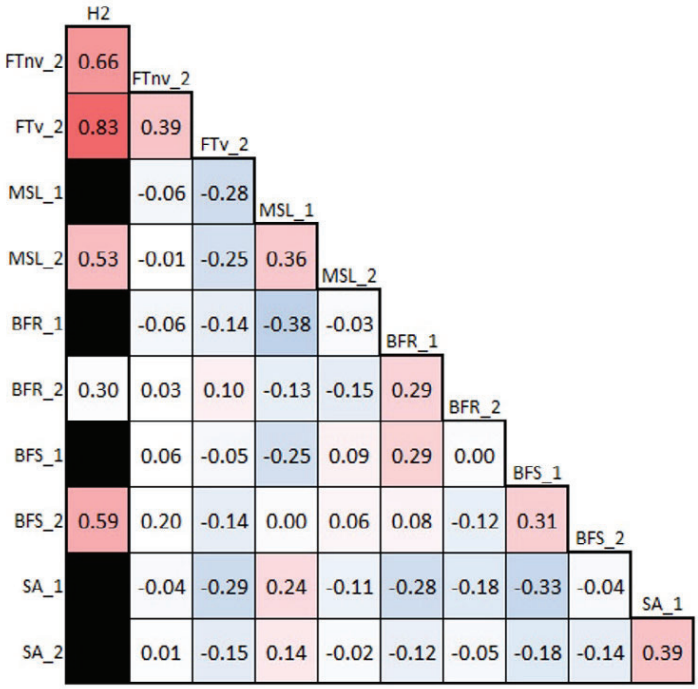


FIGURE 3 | Heritabilities of and correlation between morphological trait values. The first column lists the broad-sense heritabilities of the traits for which replicated measurements were available. The half diallel matrix lists the Pearson correlations between traits. Correlations of $r^2 < -0.15$ and $r^2 > 0.15$ were significant at $\alpha = 0.05$. The colour-scale indicates the strength of the heritability/correlation from -1 (blue) to 0 (white) to +1 (red). H2 = Broad-sense heritability; FTnv = Flowering time without vernalization (days after sowing); FTv = Flowering time after vernalization (days after transfer); MSL = Main stem length (cm); BFR = Branching from rosette (nr.); BFS = Branching from stem (nr.); SA = Seed area (mm2); The appendix _1 indicates the monoploids, while a _2 indicates the diploids.

Detection of genotype-by-ploidy interaction QTLs

Each of the four traits measured in both the mono- and diploid population were subjected to trait specific dual-trait CIM, in which measurements at the two ploidy levels were considered to be different traits. Significant QTLs could be detected for each trait in both generations. In total fifteen QTLs were detected for the various traits, for which six QTLs were detected with a significant interaction effect of the ploidy level (**Table 4**). Three genotype-by-ploidy (GxP) QTLs were detected for main stem length, while one GxP QTL was detected for each of the other traits.

For main stem length five QTLs were detected in total, with a major QTL on the top of chromosome 5 and minor QTLs on chromosomes 3 and 4 (**Table 4**). At the major QTL of chromosome 5 the Ge-0 allele increased the stem length in the monoploids (normalized effect size 0.49), whereas genotypic variation at this locus had no significant influence on the length of the diploids (effect size 0.03). The QTL on chromosome 3 showed a similar pattern with a significant effect of the Ge-0 allele in the monoploids, although with smaller effect size as the QTL on chromosome 5, but

not in the diploids. Finally, on chromosome 4, three QTLs with overlapping support intervals spanning the entire chromosome and similar effect signs were detected.

For both variation in branching from the rosette and branching from the main stem three QTLs were detected (**Table 4**). For variation in branching from the rosette, QTLs were located on the bottom of chromosomes 3 and 5 and the top of chromosome 5. The QTL on the bottom of chromosome 5 revealed a clear GxP interaction, as it was highly significant in the monoploids (p-value = <0.001) while it was not detected in the diploid generation (p-value = 0.642). The Ge-0 genotype at this QTL explained an increase in the number of branches in the monoploids, while a Ge-0 genotype at the two other QTLs decreased the number of branches from the rosette at both ploidy levels. A QTL for variation in branching from the main stem on the middle of chromosome 5 was also identified as a GxP QTL. However, here the QTL was only significantly detected in the diploids (p-value = <0.001) and not in the monoploids (p-value = 0.204). Similar to an increase in main stem length, Ge-0 alleles at any of the three detected QTLs increases the number of branches.

Finally, four QTLs were detected for variation in seed area, of which a GxP interaction was identified for the QTL on the middle of chromosome 3 (**Table 4**). This QTL was significant in the monoploids (p-value = 0.002) but not in the diploids (p-value = 0.224). However, this QTL exerted only a minor effect. Another QTL on chromosome 3 was significantly detected in both generations, although it was much weaker in the diploids (p-value = 0.043) and a large difference in the effect size of the QTL was observed (0.42 and 0.17 for monoploids and diploids, respectively). The results of the dual ploidy QTL analysis clearly indicate that differences in ploidy do not affect every genotype and trait similarly. Indeed, strong GxP QTLs explain for a large part the phenotypic differences observed between genotypes and ploidy levels.

Pleiotropic effects of genotype-by-ploidy interactions

A weak to moderate correlation could be observed between values of the different morphological traits measured in the two populations (**Fig. 3**). These relationships suggest a partial co-regulation of traits. Indeed, we detected QTLs at similar positions for multiple traits (**Table 4**). We, therefore, subjected the various traits measured in the monoploids and diploids after vernalization to a single multi-trait CIM analysis to identify possible co-location of QTLs. A total of nine QTLs were detected using this approach (**Table 5**). None of these were trait specific and only the minor QTLs on the bottom of chromosomes 1 and 4 were ploidy specific (p-value <0.01), although suggestive QTLs (p-value < 0.05) were detected for other traits or at the other ploidy level as well (**Table 5**).

TABLE 4 | Overview of the QTLs detected for phenotypic variation in the monoploid and diploid generations. Positions are shown with support intervals in between brackets. The column "*-log10(p)*" indicates the significance of the QTL for the combined ploidy levels, while in the column "*p-value*" the specific p-value for each level is given. "*Effect size*" is given as the normalized additive effect of the QTL, where positive values indicate a positive effect of the T540 allele and negative values indicate a Ge-0 high value allele; "*s.e.*" is the standard error of the mean effect; "*%EV*" is the explained variance according to a mixed model. QTLs with a significant p-value (<0.05) are indicated by bold, while when a specific ploidy level is non-significant for the QTL it is shown in grey.

Trait	Chromosome	Position (cM)	-LOG10(P)	Population	P-value	Effect size	s.e.	%EV
MSL	III	67.4 (2.9-117.2)	4.1	Diploid (DH)	0.234	0.10	0.08	0.9
				Monoploid	**<0.001**	**-0.24**	**0.07**	**5.9**
	IV	5.3 (5.3-92.8)	3.7	**Diploid (DH)**	**0.033**	**-0.15**	**0.07**	**2.3**
				Monoploid	**<0.001**	**-0.24**	**0.06**	**5.8**
	IV	57.9 (5.3-92.8)	3.3	**Diploid (DH)**	**0.030**	**-0.18**	**0.08**	**3.2**
				Monoploid	**<0.001**	**-0.26**	**0.07**	**6.9**
	IV	88.8 (5.3-92.8)	4.0	**Diploid (DH)**	**<0.001**	**-0.36**	**0.08**	**12.6**
				Monoploid	0.099	-0.12	0.07	1.3
	V	4.7 (2.4-133.1)	17.5	Diploid (DH)	0.698	0.03	0.07	0.1
				Monoploid	**<0.001**	**-0.49**	**0.06**	**23.5**
BFR	III	102.5 (2.9-117.2)	3.5	**Diploid (DH)**	**0.001**	**0.26**	**0.08**	**7**
				Monoploid	**0.009**	**0.18**	**0.07**	**3.2**
	V	4.7 (2.4-133.1)	9.3	**Diploid (DH)**	**0.006**	**0.20**	**0.08**	**4.2**
				Monoploid	**<0.001**	**0.41**	**0.07**	**17**
	V	130.5 (2.4-133.1)	6.6	Diploid (DH)	0.642	-0.03	0.07	0.1
				Monoploid	**<0.001**	**-0.35**	**0.07**	**12.5**
BFS	II	84.3 (0.9-93.5)	4.1	**Diploid (DH)**	**0.010**	**-0.17**	**0.07**	**2.9**
				Monoploid	**<0.001**	**-0.30**	**0.08**	**8.7**
	V	71 (2.4-133.1)	3.7	**Diploid (DH)**	**<0.001**	**-0.23**	**0.06**	**5.3**
				Monoploid	0.204	0.09	0.07	0.9
	V	126.4 (2.4-133.1)	14.4	**Diploid (DH)**	**<0.001**	**-0.48**	**0.06**	**23.1**
				Monoploid	**<0.001**	**-0.30**	**0.07**	**8.8**
SA	II	84.3 (0.9-93.5)	6.0	**Diploid (DH)**	**0.004**	**0.23**	**0.08**	**5.3**
				Monoploid	**<0.001**	**0.31**	**0.06**	**9.8**
	III	21.6 (2.9-117.2)	8.1	**Diploid (DH)**	**0.043**	**-0.17**	**0.08**	**2.9**
				Monoploid	**<0.001**	**-0.42**	**0.07**	**17.4**
	III	62.7 (2.9-117.2)	2.2	Diploid (DH)	0.224	-0.11	0.09	1.3
				Monoploid	**0.002**	**-0.24**	**0.08**	**5.5**
	V	91.8 (2.4-133.1)	8.0	**Diploid (DH)**	**0.006**	**-0.21**	**0.08**	**4.5**
				Monoploid	**<0.001**	**-0.37**	**0.06**	**13.4**

A minor QTL on the lower arm of chromosome 3 significantly (p-value <0.01) explained variation in all monoploid traits but only in branching from the rosette in the diploids. The T540 allele at this locus increases the number of branches from the rosette in the monoploids and diploids, even though the diploids did not display a large variation for this trait. Additionally, the same T540 allele caused an increase in branches from the stem in monoploids. However, the same allele also decreases main stem length and seed size of the monoploids. Additional minor to moderate QTLs co-locating on the lower arm of chromosomes 2 and 4 and in the middle of chromosome 3 were detected, explaining variation in multiple traits in both the mono- and diploids. The sign and effect size of these coinciding QTLs was in line with the observed correlation between these traits (**Fig. 3**).

By far the strongest and largest number of QTLs was detected on chromosome 5. Strong QTLs for variation in main stem length and rosette branching in the monoploids coincided at the top of the chromosome, although with opposite effect sign (**Table 5**). Another strong QTL for variation in the size of seeds derived from monoploids at 61.2 cM coincided with highly significant QTLs for variation in stem branching, flowering time after vernalization and main stem length of diploids. Finally, close to the end of the chromosome (121.7 cM), a strong QTL for variation in main stem length and branching of the monoploids co-located with a QTL for variation in flowering time after vernalization and branching from the rosette of diploids. The Ge-0 allele at this locus increased all trait values except flowering time after vernalization, which was delayed by the T540 allele.

Since genetic variation at the two QTLs at the top and bottom of chromosome 5 has the strongest effect on branching and main stem length (in addition to flowering time in the diploids) we analysed the effect of each of the four possible haplotypes in both the monoploid and diploid populations. Reflecting the absence of a significantly detected QTL for variation in stem length and branching at the top of chromosome 5 in the diploids, genotypic variation at the two QTLs had a much stronger effect on the monoploids (**Fig. 4**). This clearly indicates that the effect of genetic variation can be much stronger in monoploids than in diploids (**Fig. 2**).

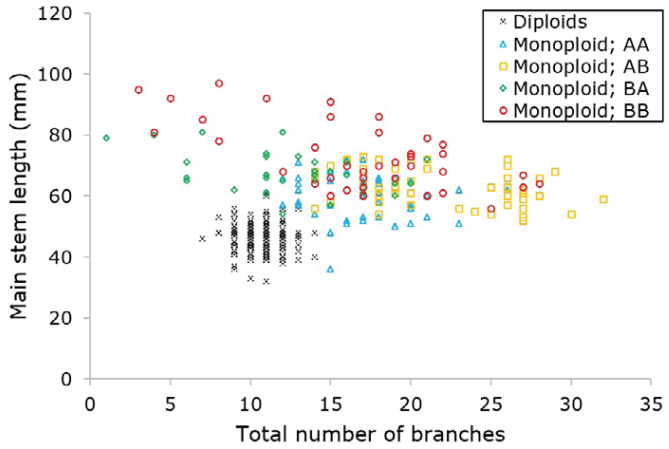


FIGURE 4 | Genotype specific phenotype response to ploidy. The total number of branches is the sum of the branches from the rosette and the main inflorescence. In the legend, the colours and indicated genotypes denote the different alleles at the QTLs at the top and bottom of chromosome 5, respectively (A = T540; B = Ge-0). The phenotypes of the diploids are depicted in black.

TABLE 5 | Overview of the QTLs detected in a single multi-trait composite interval QTL mapping. Positions are shown with support intervals in between brackets. The column "*-log10(p)*" indicates the significance of the QTL for the combined traits, while in the column "*p-value*" the specific p-value for each trait is given. "*Effect size*" is given as the normalized additive effect of the QTL, where positive values indicate a positive effect of the T540 allele and negative values indicate a Ge-0 high value allele; "*s.e.*" is the standard error of the mean effect; "*%EV*" is the explained variance according to a mixed model. Traits with a non-significant p-value (>0.01) for the QTL are in grey. Traits with a p-value between 0.05 and 0.01 for the QTL are in grey and bold. Traits with a highly significant p-value (<0.01) for the QTL are in bold and black.

Chromosome	Position (cM)	-LOG10(P)	Trait	Population	P-value	Effect size	s.e.	%EV
I	136.4 (5.2-150.6)	8.5	**FT_v**	**Diploid (DH)**	**<0.001**	**0.19**	**0.05**	**3.7**
			MSL	Diploid (DH)	0.356	0.06	0.07	0.4
			MSL	**Monoploid**	**0.023**	**0.13**	**0.06**	**1.7**
			BFR	**Diploid (DH)**	**0.009**	**0.19**	**0.07**	**3.5**
			BFR	**Monoploid**	**0.012**	**0.16**	**0.07**	**2.6**
			BFS	Diploid (DH)	0.434	-0.05	0.07	0.3
			BFS	**Monoploid**	**0.016**	**-0.17**	**0.07**	**2.8**
			SA	**Diploid (DH)**	**0.028**	**-0.16**	**0.07**	**2.6**
			SA	**Monoploid**	**0.030**	**-0.13**	**0.06**	**1.7**
II	80.4 (0.9-93.5)	6.3	**FT_v**	**Diploid (DH)**	**<0.001**	**-0.20**	**0.05**	**3.9**
			MSL	**Diploid (DH)**	**0.011**	**-0.17**	**0.07**	**3**
			MSL	Monoploid	0.148	-0.09	0.06	0.8
			BFR	Diploid (DH)	0.157	0.11	0.08	1.1
			BFR	Monoploid	0.541	-0.04	0.07	0.2
			BFS	Diploid (DH)	0.071	-0.12	0.07	1.5
			BFS	**Monoploid**	**0.018**	**-0.17**	**0.07**	**2.9**
			SA	**Diploid (DH)**	**0.013**	**0.19**	**0.08**	**3.7**
			SA	**Monoploid**	**<0.001**	**0.26**	**0.06**	**6.5**

Chromosome	Position (cM)	-LOG10(P)	Trait	Population	P-value	Effect size	s.e.	%EV
III	21.6 (2.9-117.2)	11.7	**FT_v**	**Diploid (DH)**	**<0.001**	**0.19**	**0.05**	**3.7**
			MSL	Diploid (DH)	0.540	0.04	0.07	0.2
			MSL	Monoploid	0.907	0.01	0.06	0
			BFR	**Diploid (DH)**	**0.037**	**0.16**	**0.08**	**2.6**
			BFR	Monoploid	0.258	0.08	0.07	0.6
			BFS	Diploid (DH)	0.091	0.12	0.07	1.4
			BFS	**Monoploid**	**0.025**	**0.17**	**0.07**	**2.8**
			SA	**Diploid (DH)**	**0.017**	**-0.19**	**0.08**	**3.6**
			SA	**Monoploid**	**<0.001**	**-0.49**	**0.07**	**24.2**
III	92.9 (2.9-117.2)	6.1	**FT_v**	**Diploid (DH)**	**0.036**	**0.10**	**0.05**	**1.1**
			MSL	Diploid (DH)	0.338	-0.06	0.07	0.4
			MSL	**Monoploid**	**0.002**	**-0.18**	**0.06**	**3.1**
			BFR	**Diploid (DH)**	**0.001**	**0.25**	**0.07**	**6**
			BFR	**Monoploid**	**0.009**	**0.17**	**0.07**	**2.9**
			BFS	Diploid (DH)	0.348	-0.06	0.07	0.4
			BFS	**Monoploid**	**0.003**	**0.20**	**0.07**	**4**
			SA	Diploid (DH)	0.345	-0.07	0.07	0.5
			SA	**Monoploid**	**<0.001**	**-0.26**	**0.06**	**6.6**
IV	51.9 (5.3-92.8)	8.5	**FT_v**	**Diploid (DH)**	**<0.001**	**0.23**	**0.05**	**5.4**
			MSL	**Diploid (DH)**	**<0.001**	**-0.32**	**0.07**	**10.1**
			MSL	**Monoploid**	**<0.001**	**-0.29**	**0.06**	**8.2**
			BFR	Diploid (DH)	0.203	0.10	0.08	1
			BFR	Monoploid	0.088	0.12	0.07	1.5
			BFS	Diploid (DH)	0.653	-0.03	0.07	0.1
			BFS	**Monoploid**	**0.042**	**0.15**	**0.08**	**2.4**
			SA	**Diploid (DH)**	**0.014**	**-0.20**	**0.08**	**3.9**
			SA	Monoploid	0.051	-0.13	0.07	1.6
IV	88.8 (5.3-92.8)	3.7	**FT_v**	**Diploid (DH)**	**0.003**	**0.17**	**0.06**	**3**
			MSL	**Diploid (DH)**	**<0.001**	**-0.34**	**0.08**	**11.7**
			MSL	Monoploid	0.149	-0.10	0.07	1
			BFR	Diploid (DH)	0.796	0.02	0.09	0
			BFR	Monoploid	0.709	-0.03	0.08	0.1
			BFS	Diploid (DH)	0.931	-0.01	0.08	0
			BFS	Monoploid	0.536	0.05	0.08	0.3
			SA	Diploid (DH)	0.269	0.10	0.09	0.9
			SA	Monoploid	0.550	0.04	0.07	0.2

Chromosome	Position (cM)	-LOG10(P)	Trait	Population	P-value	Effect size	s.e.	%EV
V	4.7 (2.4-133.1)	15.7	FT_v	Diploid (DH)	0.190	-0.07	0.05	0.4
			MSL	Diploid (DH)	0.973	0.00	0.07	0
			MSL	**Monoploid**	**<0.001**	**-0.50**	**0.06**	**25.4**
			BFR	**Diploid (DH)**	**0.009**	**0.20**	**0.08**	**3.9**
			BFR	**Monoploid**	**<0.001**	**0.37**	**0.07**	**13.7**
			BFS	Diploid (DH)	0.136	0.10	0.07	1
			BFS	**Monoploid**	**0.039**	**0.15**	**0.07**	**2.1**
			SA	Diploid (DH)	0.445	-0.06	0.08	0.3
			SA	Monoploid	0.068	-0.11	0.06	1.3
V	61.2 (2.4-133.1)	10.8	**FT_v**	**Diploid (DH)**	**<0.001**	**-0.24**	**0.05**	**5.8**
			MSL	**Diploid (DH)**	**0.001**	**0.24**	**0.07**	**5.6**
			MSL	**Monoploid**	**0.016**	**0.15**	**0.06**	**2.3**
			BFR	Diploid (DH)	0.859	0.01	0.08	0
			BFR	Monoploid	0.425	0.06	0.07	0.3
			BFS	**Diploid (DH)**	**<0.001**	**-0.25**	**0.07**	**6.4**
			BFS	Monoploid	0.561	0.04	0.08	0.2
			SA	**Diploid (DH)**	**0.018**	**-0.19**	**0.08**	**3.6**
			SA	**Monoploid**	**<0.001**	**-0.24**	**0.07**	**5.9**
V	121.7 (66.6-133.1)	77.6	**FT_v**	**Diploid (DH)**	**<0.001**	**0.64**	**0.05**	**40.4**
			MSL	Diploid (DH)	0.671	-0.03	0.07	0.1
			MSL	**Monoploid**	**<0.001**	**-0.24**	**0.06**	**5.9**
			BFR	Diploid (DH)	0.474	-0.05	0.08	0.3
			BFR	**Monoploid**	**<0.001**	**-0.38**	**0.07**	**14.3**
			BFS	**Diploid (DH)**	**<0.001**	**-0.45**	**0.07**	**20.2**
			BFS	**Monoploid**	**<0.001**	**-0.32**	**0.07**	**10.3**
			SA	Diploid (DH)	0.715	-0.03	0.08	0.1
			SA	Monoploid	0.159	-0.09	0.06	0.8

DISCUSSION

Application of a late flowering doubled haploid mapping population

It is well known that different accessions of Arabidopsis respond differently to environmental conditions (*24, 26*). For instance, day-length sensitivity and vernalization requirement determine for a large part the discrimination between winter- and summer-annuals (*27, 28*). Moreover, when mapping populations are subjected to short or long day-length conditions with or without vernalization, differences in the number and strength of detected flowering time QTLs can be observed (*29*). The use of a haploid-inducer line in this study allowed the generation of a homozygous mapping population from underexploited late flowering accessions (*13*). As such, a diploid population could be developed in only three generations. For this population, QTL mapping for variation in flowering time in two different environments (i.e. with and without vernalization) was performed.

In addition to a number of minor QTLs, a major QTL for variation in flowering time of vernalized plants was detected near the previously described and identified *VERNALIZATION INSENSITIVE 3* (*VIN3*; At5g57830) locus at the bottom of chromosome 5 (*29-31*). Previously, variation in flowering time associated with this locus was explained by an indel of three nucleotides within an exon of *VIN3* (*31*). According to the SALK Arabidopsis genome browser (http://signal.salk.edu/) this indel is not polymorphic for Ge-0 and T540, although multiple other SNPs differentiate the intronic and promotor region of *VIN3* of these accessions, including 28 nucleotides, which seem to be deleted from the T540 *VIN3* promotor compared to Ge-0 (**Fig. S4**).

A second gene, *REDUCED VERNALIZATION RESPONSE 2* (*VRN2*; At4g16845), related to response to vernalization was located within the support interval of a QTL for variation in flowering time after vernalization, located on chromosome 4 (*32*). The *VRN2* protein mediates vernalization through interaction with the Polycomb Group (PcG) protein complex including *VIN3* (*33, 34*). This PcG complex is known to interact with, and cause the stable reduction of the expression levels of, the floral repressor *FLOWERING LOCUS C* (*FLC*; At5g10140) (*33, 34*), which collocates with the position of a flowering time QTL on the top of chromosome 5. This QTL was also detected for variation in main stem length, which strongly suggests a pleiotropic effect on the inflorescence architecture and flowering pathways, previously attributed to *FLC* (*23*).

The detection of flowering time QTLs in a segregating mapping population of late flowering accessions, especially after vernalization, clearly identifies major QTLs other than those usually associated with flowering time variation in early accessions. This suggests that the regulation of flowering time in late accessions is controlled by

variation at other loci than those in early flowering accessions (*e.g*. *FRI* and *FLC*). It is likely that flowering time is not the only trait that discriminates summer annuals from winter annuals, which advocates the analysis of traits in late flowering populations in addition to the abundantly available early flowering populations.

Effects of haploidization on phenotypic variation

Exploiting the availability of a mono- and diploid Arabidopsis mapping population, QTL analyses were applied to map and compare possible ploidy-dependent effects. A multi-trait CIM analysis resulted in the detection of six QTLs with a GxP interaction, while additional QTLs showed large differences in effect sizes at either ploidy level. An obvious explanation for the GxP QTLs is that monoploid plants are sterile due to unbalanced segregation of the chromosomes during meiosis. Indeed, although not explicitly quantified, monoploids displayed an extended period of flowering compared to fertile diploids, possibly causing the increase in main stem length. Similarly, the development of exceptionally high numbers of rosette branches (*35*) increased the total number of flowers produced. This suggests that the plants attempt to compensate for the lack of viable seed production by an increase in reproductive tissue formation, implying that the QTLs detected specifically for monoploids might be involved in the response to sterility. A similar phenomenon of additional branch formation has been described for the male sterile Landsberg *erecta* mutant (*ms1*-L*er*) (*36*). It remains interesting that QTLs explaining this variation could be detected, as this indicates that these two accessions respond differently to haploidization.

Based on the antagonistic effect of the QTL on the top of chromosome 5 for either additional rosette branch formation (inferred by the T540 allele) or taller growth (inferred by the Ge-0 allele), it is strongly suggested that both accessions follow a different morphological approach to achieve a similar increase in the number of flowers. The fact that a single QTL is identified for variation in rosette branching and main stem length could be due to one of the many pleiotropic genes that function in the control of inflorescence architecture (*37*).

Possible candidate genes may be part of the florigen gene family (*38*) which is known to function as a mobile flowering time switch. For instance, *FLOWERING LOCUS T* (*FT*; At1g65480) and *TWIN SISTER OF FT* (*TSF*; AT4G20370), are known to function in both flower induction and shoot branching pathways (*39*). Another member of the same gene family, *TERMINAL FLOWERING 1* (*TFL1*; At5g03840), is located within the support interval of the QTL at the top of chromosome 5 and has been shown to be involved in flowering architecture (*40*). Although no variation within the *TFL1* coding sequence could be observed between the two accessions, according to the 1001 genome sequence browser (http://signal.salk.edu/), several SNPs and possibly

deletions within the promotor region of the T540 allele might cause a differential expression of this gene (**Fig. S5A**).

Assuming that flowering architecture is not influenced by *VIN3*, an alternative candidate explaining the effect of the QTL at the bottom of chromosome 5 on both branching and main stem length is *AUXIN RESPONSE FACTOR2* (*ARF2*; At5g62000), which is involved in multiple developmental processes via cell proliferation (*41, 42*). Again, sequence-based evidence suggests that T540 and Ge-0 possess functionally different alleles (**Fig. S5B**). Moreover, a knock-down of *ARF2* leads to an increase in stem length and a sterile phenotype (*41*).

Additional QTLs with likely candidates are for instance the QTL explaining variation in the size of seeds on the top of chromosome 3. This QTL has been identified previously as *HAIKU 2* (*IKU2*; At3g19700) and this gene is a likely candidate since it functions in the endosperm growth pathway (*43*). In addition, a monoploid specific QTL on chromosome 2 explaining variation in branching from the stem coincides with the previously identified *AGAMOUS-LIKE 6* gene (*AG6*; a.k.a. *REDUCED SHOOT BRANCHING 1*; AT2G45650), to which pleiotropic phenotypic effects on both the flowering and branching pathways have been previously attributed (*44*).

Although sterility might be causal for some of the GxP interactions of the QTLs, it is possible that other molecular processes are of influence as well. In previous studies on ploidy series including monoploids, performed in maize (*8*), yeast (*9*), potato (*10*) and Chinese cabbage (*11*), differentially expressed genes were identified at different ploidy levels, indicating a specific sensitivity to ploidy, instead of sterility. Moreover, in a dosage series (*x* , 2*x*, 4*x*) of maize inbred lines (*45*), genetic background and ploidy was suggested to interact. Further evidence for GxP interactions independent of sterility come from an RNA-seq comparison of diploid and tetraploid Arabidopsis accessions, in which Col-0 and L*er*-0 displayed different numbers of upregulated genes at the tetraploid level (*46*). In both studies it was argued that the altered nuclear surface to volume ratio might have caused the differential expression of genes. However, clear mechanisms explaining how these altered ratio's cause gene expression differences are so far elusive. Despite the uncertainty of the possible mechanisms of GxP interactions it is clear that the mapping of quantitative traits in mono- and diploids can reveal additional variation, which might be instrumental in the elucidation of the genetic regulation of complex traits.

EXPERIMENTAL PROCEDURES

Population development

Two late flowering accessions, T540 (CS76239) from Sweden and Ge-0 (CS76135) from Switzerland were selected based on phenotypic differences and expected unexplored genotypic differences compared to widely used early flowering accessions. These accessions were crossed to produce a biparental hybrid F_1. The F_1 (T540 x Ge-0) was used as a pollen donor and crossed to the GFP-tailswap haploid-inducer line to generate monoploid offspring (*12*). From these crosses, 250 viable seeds were sown and 210 putative monoploid lines were selected based on morphology during growth (*14*). Spontaneous genome doubling in the monoploids followed by selfing created a set of 171 unique diploid homozygous lines.

Plant growth conditions

All seeds from a cross between the F_1 hybrid (T540 x Ge-0) and the GFP-tailswap line were sown on ½ MS agar plates without sucrose. The seeds on these plates were stratified for four days at 4°C in darkness and subsequently placed in a climate chamber at 25°C with a diurnal cycle of 16 hours of light and 8 hours of darkness to induce seed germination. After two days of pre-germination, only potential monoploid seedlings were transplanted to wet Rockwool blocks of 4 x 4 cm in a climate chamber (16h LD, 125 µmolm-2s-1, 70% RH, 20/18˚C day/night cycle). All plants were watered three days per week for 5 min with 1/1000 Hyponex solution (Hyponex, Osaka, Japan) using flooding tables. Here they remained for three weeks to allow growth before vernalization. Vernalization was performed for eight weeks (12h LD, 125 µmolm-2s-1, 70% RH, 4˚C constant). After vernalization, plants were transferred to the greenhouse where they were allowed to flower and mature. Monoploid plants were selected based on morphology as described before (*14*). Subsequently, diploid seeds were harvested after recording phenotypic traits of the monoploids.

The second experiment included ten replicates for each of 210 assumed diploids. These were stratified on wet filter paper in similar conditions as the agar plates of the previous experiment. Subsequently, five of the seedlings were grown similar to the monoploids, including three weeks growth in long day conditions and vernalization for eight weeks, while the five other replicates were transferred to the greenhouse. The five replicates in the greenhouse conditions were allowed to grow in a completely randomized design without vernalization for a maximum of 100 days after transfer or until flowering or senescence. The five diploids that underwent vernalization

remained in climate chambers with similar conditions as pre-vernalization (16h LD, 125 µmolm-2s-1, 70% RH, 20/18˚C day/night cycle). The plants were randomized in a completely randomized design where they were allowed to grow for a maximum of ninety days.

Phenotypic measurements

The monoploids were phenotyped for the number of branches from the rosette and branching from the stem, main stem length (cm) and seed area (approximately 100 seeds were taken three times from the same storage bag for three separate photos, these were analysed for seed area) during harvesting. For the second experiment, the same four phenotypes were measured. However, now also flowering time before and after vernalization was included as a phenotype. Flowering time without vernalization was measured as the number of days after planting until the first flower on the main stem opened its petals. Flowering time with vernalization was measured as the number of days after vernalization until the first flower on the main stem opened its petals. Plants that did not germinate or that died within the period of the experiment were discarded. For the plants used for genotyping only flowering time was recorded, as taking a flower head, used for extracting DNA, from the plant might influence the other traits. All the monoploid and F_2 phenotypes are based on a single observation per genotype, while for the DH population, which were measured with five replicates, the reported values are the means.

Genotyping of the populations

For 210 doubled haploids and 71 F_2s the DNA was extracted from flower heads by applying a CTAB DNA extraction protocol which was adapted for use on 96 well plates. Genotyping was performed using a GoldenGate Assay from Illumina, using 384 SNP markers. Of those, 142 markers were polymorphic for the two parental lines. Of these 142, only 114 markers showed nonredundant recombination patterns for either the diploids or F_2s. Nine additional KASP markers were included to a total of 123 markers (**Table S1**). From 210 selected diploid lines, 195 were successfully genotyped and only four were discarded because of too much heterozygosity or missing data. Eventually, only 171 diploids were used for the final analyses because of redundant genotypes and lack of data in either mono- or diploid generation.

Genetic map comparison of the doubled haploid population and F_2 population

To confirm no anomalies were present in the doubled haploids (DHs), a comparison with an F_2 population was performed. Individual lines from both populations were genotyped and genetic maps were generated. A subset of 71 F_2s and 171 DHs were successfully genotyped. Genetic maps were constructed for both the F_2 and the DHs independently using Kosambi's regression mapping function in JoinMap 6.1 (**Fig. S1**). Segregation distortions were determined by GenStat 19 edition (**Fig. S2**). The DH map was also used for the genetic mapping in monoploids.

Statistical analyses and QTL mapping

Pearson correlations between traits were calculated using the cor function in R. The broad-sense heritabilities of the doubled haploids were calculated in R using the repeatability function of the heritability package (*47*). QTL analyses were performed using GenStat (19th edition)(*48*), where mean phenotypic values per DH line were used and single observations in the case of the monoploids and F_2s. In order to have a maximum QTL effect and QTL x E or QTL x Ploidy interaction detection, we first analysed the separate traits using single-trait multiple environment composite interval mapping (where either vernalization or the ploidy level was considered as the environment). The final analyses encompassed a multi-trait single environment analyses, including all traits measured after vernalization. First an initial analysis of simple interval mapping was performed with a maximum step size of 5 cM along the genome. Other settings were kept as default (maximum cofactor proximity = 50 cM; minimum distance for QTL selection = 30 cM; threshold for genome-wide significance level = α = 0.05). After these first analyses, markers associated with candidate QTLs were automatically set as cofactors for the composite interval mapping. The QTLs that resulted from this scan were tested for interaction effects in the selection of a final QTL model.

ACKNOWLEDGEMENTS

We like to express our gratitude to G. Stunnenberg and T. Stoker of Wageningen University for technical assistance during experimental work and E. Kerdaffrec, M. Nordborg, of the Gregor Mendel Institute in Vienna, Austria for performing the golden-gate genotyping assays.

REFERENCES

1. D.-Y. Chao *et al.*, Polyploids Exhibit Higher Potassium Uptake and Salinity Tolerance in Arabidopsis. *Science* 341, 658-659 (2013).
2. J. C. Del Pozo, E. Ramirez-Parra, Deciphering the molecular bases for drought tolerance in Arabidopsis autotetraploids. *Plant, Cell & Environment* 37, 2722-2737 (2014).
3. A. Fort *et al.*, Disaggregating polyploidy, parental genome dosage and hybridity contributions to heterosis in Arabidopsis thaliana. *New Phytol* 209, 590-599 (2016).
4. I. M. Henry *et al.*, Aneuploidy and Genetic Variation in the Arabidopsis thaliana Triploid Response. *Genetics* 170, 1979-1988 (2005).
5. M. Orzechowska, S. Gurdek, D. Siwinska, A. Piekarska-Stachowiak, Cytogenetic characterization of the Arabidopsis thaliana natural tetraploid ecotype Warschau stability during in vitro regeneration. *Plant Cell, Tissue and Organ Culture (PCTOC)* 126, 553-560 (2016).
6. Z. Yu, K. Haage, V. E. Streit, A. Gierl, R. A. Torres Ruiz, A large number of tetraploid Arabidopsis thaliana lines, generated by a rapid strategy, reveal high stability of neo-tetraploids during consecutive generations. *Theoretical and Applied Genetics* 118, 1107-1119 (2009).
7. C. L. Schiff, I. W. Wilson, S. C. Somerville, Polygenic powdery mildew disease resistance in Arabidopsis thaliana: quantitative trait analysis of the accession Warschau-1. *Plant Pathology* 50, 690-701 (2001).
8. M. Guo, D. Davis, J. A. Birchler, Dosage Effects on Gene Expression in a Maize Ploidy Series. *Genetics* 142, 1349-1355 (1996).
9. T. Galitski, A. J. Saldanha, C. A. Styles, E. S. Lander, G. R. Fink, Ploidy Regulation of Gene Expression. *Science* 285, 251-254 (1999).
10. R. M. Stupar *et al.*, Phenotypic and Transcriptomic Changes Associated With Potato Autopolyploidization. *Genetics* 176, 2055-2067 (2007).
11. A. X. Gu *et al.*, Analyses of phenotype and ARGOS and ASY1 expression in a ploidy Chinese cabbage series derived from one haploid. *Breeding Science* 66, 161-168 (2016).
12. M. Ravi, S. W. L. Chan, Haploid plants produced by centromere-mediated genome elimination. *Nature* 464, 615-618 (2010).
13. D. K. Seymour *et al.*, Rapid creation of Arabidopsis doubled haploid lines for quantitative trait locus mapping. *Proceedings of the National Academy of Sciences* 109, 4227-4232 (2012).
14. E. Wijnker *et al.*, Hybrid recreation by reverse breeding in Arabidopsis thaliana. *Nat. Protocols* 9, 761-772 (2014).
15. J. F. Crow, Haldane, Bailey, Taylor and recombinant-inbred lines. *Genetics* 176, 729-732 (2007).
16. C. Lister, C. Dean, Recombinant inbred lines for mapping RFLP and phenotypic markers in Arabidopsis thaliana. *Plant J* 4, 745-750 (1993).
17. C. Alonso-Blanco *et al.*, Development of an AFLP based linkage map of Ler, Col and Cvi Arabidopsis thaliana ecotypes and construction of a Ler/Cvi recombinant inbred line population. *Plant J* 14, 259-271 (1998).
18. M. E. El-Lithy *et al.*, New Arabidopsis Recombinant Inbred Line Populations Genotyped Using SNPWave and Their Use for Mapping Flowering-Time Quantitative Trait Loci. *Genetics* 172, 1867-1876 (2006).
19. M. Simon *et al.*, Quantitative Trait Loci Mapping in Five New Large Recombinant Inbred Line Populations of Arabidopsis thaliana Genotyped With Consensus Single-Nucleotide Polymorphism Markers. *Genetics* 178, 2253-2264 (2008).
20. C. M. O'Neill *et al.*, Six new recombinant inbred populations for the study of quantitative traits in Arabidopsis thaliana. *Theoretical and Applied Genetics* 116, 623-634 (2008).
21. E. Wijnker *et al.*, Reverse breeding in Arabidopsis thaliana generates homozygous parental lines from a heterozygous plant. *Nat Genet* 44, 467-470 (2012).
22. N. Fulcher, K. Riha, Using Centromere Mediated Genome Elimination to Elucidate the Functional Redundancy of Candidate Telomere Binding Proteins in Arabidopsis thaliana. *Frontiers in Genetics* 6, (2016).
23. X. Huang, J. Ding, S. Effgen, F. Turck, M. Koornneef, Multiple loci and genetic interactions involving flowering time genes regulate stem branching among natural variants of Arabidopsis. *New Phytol* 199, 843-857 (2013).
24. J. Lempe *et al.*, Diversity of flowering responses in wild Arabidopsis thaliana strains. *PLoS Genet* 1, 109-118 (2005).

25. M. C. Ungerer, S. S. Halldorsdottir, J. L. Modliszewski, T. F. C. Mackay, M. D. Purugganan, Quantitative trait loci for inflorescence development in Arabidopsis thaliana. *Genetics* 160, 1133-1151 (2002).
26. M. Koornneef, C. Alonso-Blanco, D. Vreugdenhil, Naturally occurring genetic variation in *Arabidopsis thaliana*. *Annual Review of Plant Biology* 55, 141-172 (2004).
27. M. Romera-Branchat, F. Andres, G. Coupland, Flowering responses to seasonal cues: what's new? *Curr Opin Plant Biol* 21, 120-127 (2014).
28. F. Andres, G. Coupland, The genetic basis of flowering responses to seasonal cues. *Nat Rev Genet* 13, 627-639 (2012).
29. C. Alonso-Blanco, S. E. El-Assal, G. Coupland, M. Koornneef, Analysis of natural allelic variation at flowering time loci in the Landsberg erecta and Cape Verde Islands ecotypes of Arabidopsis thaliana. *Genetics* 149, 749-764 (1998).
30. E. L. Dittmar, C. G. Oakley, J. Ågren, D. W. Schemske, Flowering time QTL in natural populations of Arabidopsis thaliana and implications for their adaptive value. *Molecular Ecology* 23, 4291-4303 (2014).
31. M. A. Grillo, C. Li, M. Hammond, L. Wang, D. W. Schemske, Genetic architecture of flowering time differentiation between locally adapted populations of Arabidopsis thaliana. *New Phytol* 197, 1321-1331 (2013).
32. A. R. Gendall, Y. Y. Levy, A. Wilson, C. Dean, The VERNALIZATION 2 Gene Mediates the Epigenetic Regulation of Vernalization in Arabidopsis. *Cell* 107, 525-535 (2001).
33. R. Bastow *et al.*, Vernalization requires epigenetic silencing of FLC by histone methylation. *Nature* 427, 164-167 (2004).
34. S. Sung, R. M. Amasino, Vernalization in Arabidopsis thaliana is mediated by the PHD finger protein VIN3. *Nature* 427, 159-164 (2004).
35. I. M. Ehrenreich, P. A. Stafford, M. D. Purugganan, The Genetic Architecture of Shoot Branching in *Arabidopsis thaliana*: A Comparative Assessment of Candidate Gene Associations vs Quantitative Trait Locus Mapping. *Genetics* 176, 1223-1236 (2007).
36. L. L. Hensel, M. A. Nelson, T. A. Richmond, A. B. Bleecker, The Fate of Inflorescence Meristems Is Controlled by Developing Fruits in Arabidopsis. *Plant Physiology* 106, 863-876 (1994).
37. C. Rameau *et al.*, Multiple pathways regulate shoot branching. *Frontiers in Plant Science* 5, (2015).
38. Daniel P. Wickland, Y. Hanzawa, The FLOWERING LOCUS T/TERMINAL FLOWER 1 Gene Family: Functional Evolution and Molecular Mechanisms. *Molecular Plant* 8, 983-997 (2015).
39. K. Hiraoka, A. Yamaguchi, M. Abe, T. Araki, The Florigen Genes FT and TSF Modulate Lateral Shoot Outgrowth in Arabidopsis thaliana. *Plant and Cell Physiology* 54, 352-368 (2012).
40. K. Baumann *et al.*, Changing the spatial pattern of TFL1 expression reveals its key role in the shoot meristem in controlling Arabidopsis flowering architecture. *Journal of Experimental Botany* 66, 4769-4780 (2015).
41. Y. Okushima, I. Mitina, H. L. Quach, A. Theologis, AUXIN RESPONSE FACTOR 2 (ARF2): a pleiotropic developmental regulator. *Plant J* 43, 29-46 (2005).
42. M. C. Schruff *et al.*, The AUXIN RESPONSE FACTOR 2 gene of *Arabidopsis* links auxin signalling, cell division, and the size of seeds and other organs. *Development* 133, 251-261 (2006).
43. M. Luo, E. S. Dennis, F. Berger, W. J. Peacock, A. Chaudhury, MINISEED3 (MINI3), a WRKY family gene, and HAIKU2 (IKU2), a leucine-rich repeat (LRR) KINASE gene, are regulators of seed size in Arabidopsis. *Proceedings of the National Academy of Sciences of the United States of America* 102, 17531-17536 (2005).
44. X. Huang, S. Effgen, R. C. Meyer, K. Theres, M. Koornneef, Epistatic Natural Allelic Variation Reveals a Function of AGAMOUS-LIKE6 in Axillary Bud Formation in Arabidopsis. *The Plant Cell* 24, 2364-2379 (2012).
45. N. C. Riddle, A. Kato, J. A. Birchler, Genetic variation for the response to ploidy change in Zea mays L. *Theoretical and Applied Genetics* 114, 101-111 (2006).
46. Z. Yu *et al.*, Impact of natural genetic variation on the transcriptome of autotetraploid Arabidopsis thaliana. *Proceedings of the National Academy of Sciences* 107, 17809-17814 (2010).
47. W. Kruijer *et al.*, Marker-based estimation of heritability in immortal populations. *Genetics* 199, 379-398 (2015).
48. M. Boer *et al.*, *A Guide to QTL Analysis in Genstat*. A Guide to QTL Analysis in Genstat (VSN International Hertfordshire, UK, 2015).

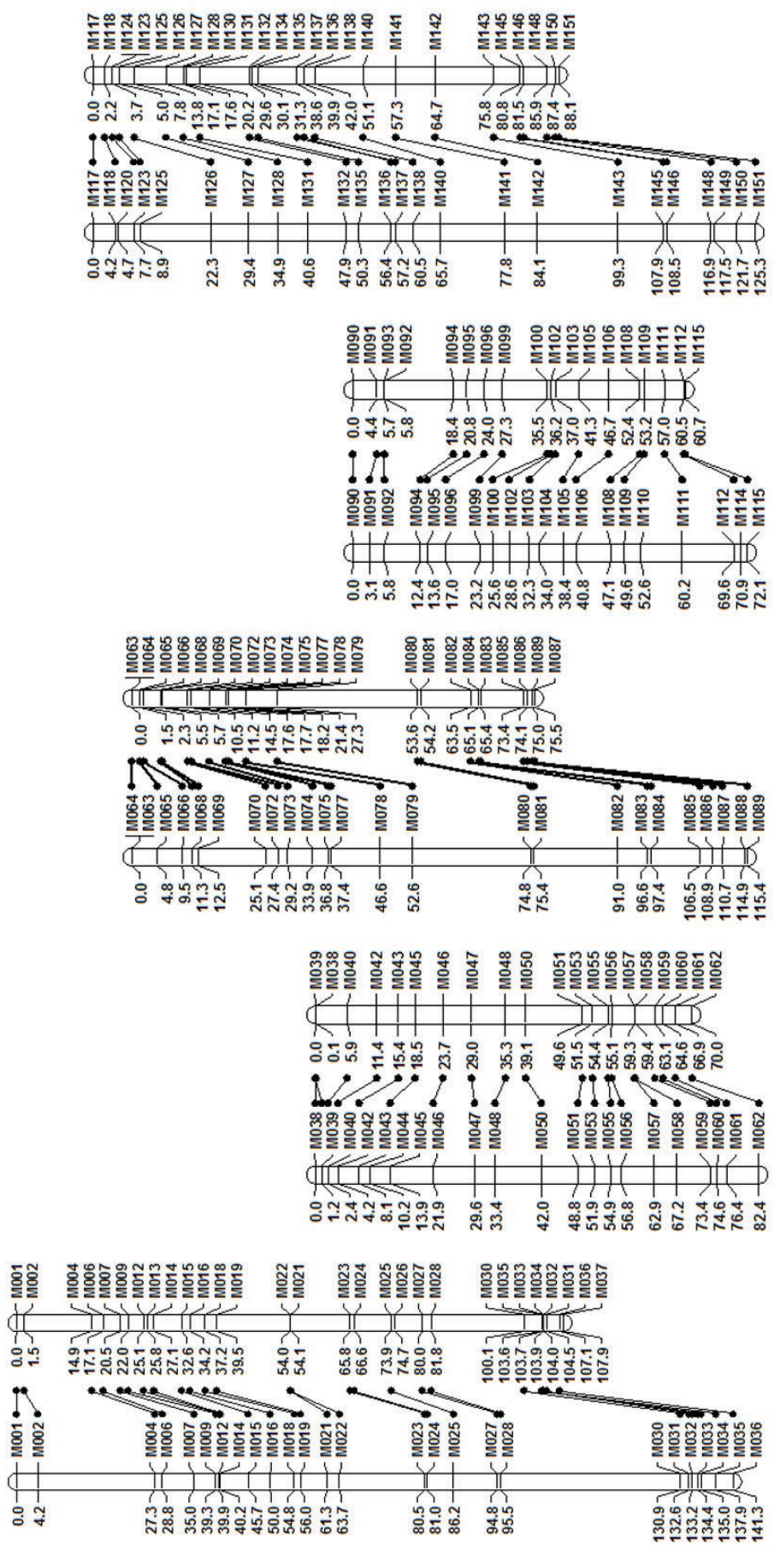


FIGURE S1 | A comparison of the genetic maps of the doubled haploids and F_2 populations. The doubled haploid maps are on the left, while the F_2 map is on the right. On the left of each map is the genetic distance in cM, while the marker names are on the right.

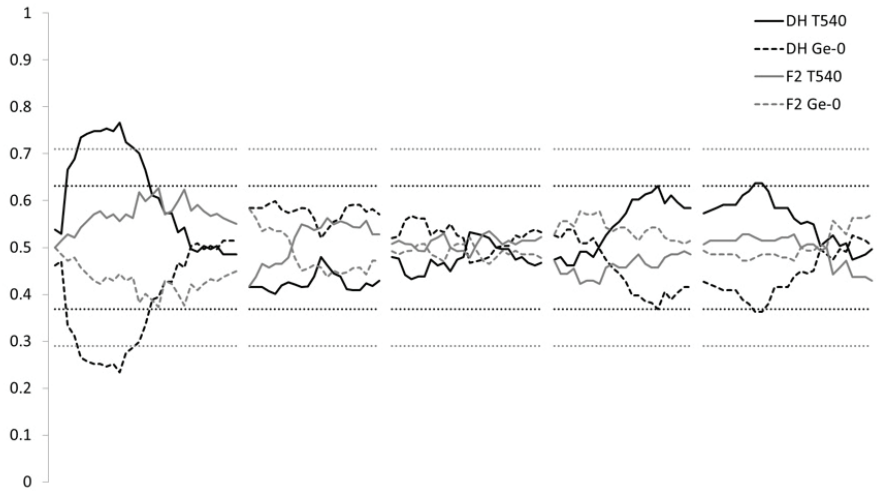


FIGURE S2 | Allele frequencies of doubled haploid and F2 populations. In black (doubled haploid) and grey (F_2) are the frequencies of the T540 (solid) and Ge-0 (dashed) alleles plotted per chromosome. The horizontal dashed lines shows the threshold for the critical value of the X^2-distribution for the doubled haploid population.

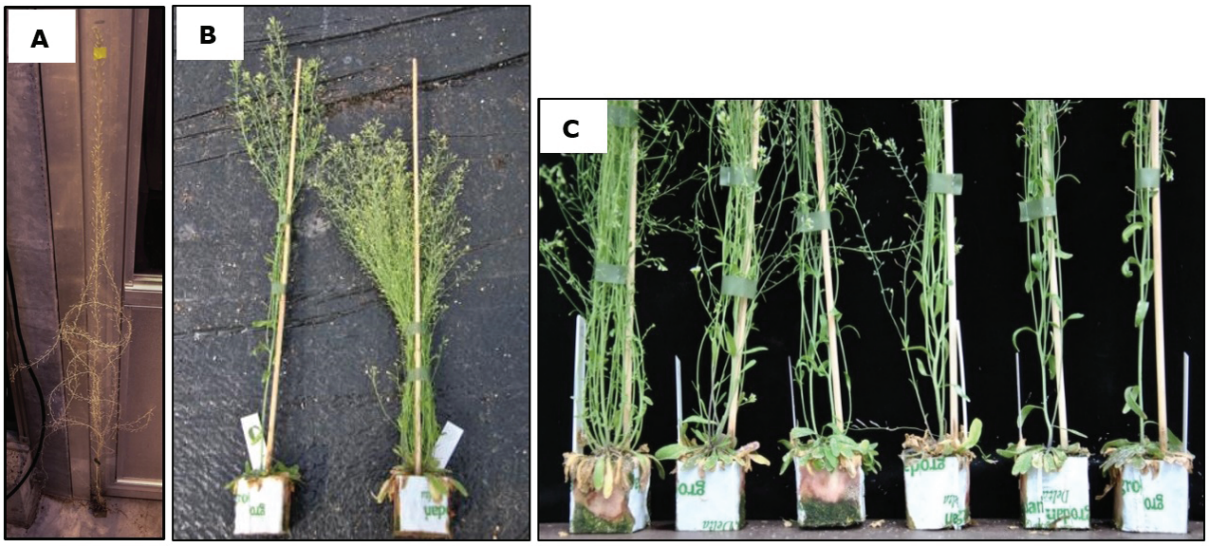


FIGURE S3 | Example of extreme monoploid plants. A) A single monoploid plant derived from a Col-0 x L*er* F1. This plant is shown to represents the extreme plant height monoploids can reach. This specific monoploid stands in a Rockwool block of 4 x 4 cm and has a total length of 143 cm. B-C) A set of monoploids derived from T-540 x Ge-0 that show variation in the shoot architecture. Note the difference in length (B) and the different number of shoot branches (C).

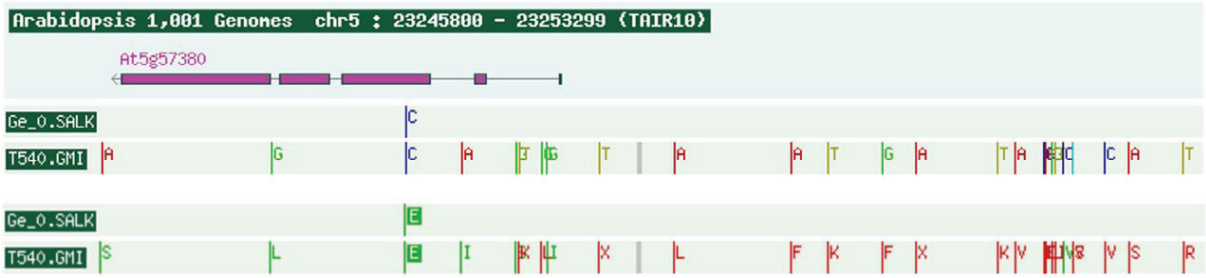


FIGURE S4 | Genomic sequence variation at the VIN3 locus. The genomic sequences of Ge-0 and T540 are represented in both their DNA sequence variation (top) and amino acid sequence variation (bottom). Many SNPs are present between the two sequences that result in non-synonymous amino acid substitutions (indicated in red at the bottom panel) in a 3kb upstream region.

FIGURE S5 | Genomic sequence variation at the TFL1 and ARF2 loci. The genomic sequences for different loci between the accessions T-540 and Ge-0 at both the DNA sequence level and amino acid level. A) The TFL1 locus is a possible candidate for the locus on the top of chromosome 5. B) The ARF2 locus is a candidate gene for the locus at the bottom of chromosome 5.

Chapter 3

A complete chromosome substitution mapping panel reveals genome-wide epistasis in Arabidopsis

Cris L. Wijnen[1†], Ramon Botet[1†], José van de Belt[1], Laurens Deurhof[1], Hans de Jong[1], C. Bastiaan de Snoo[2], Rob Dirks[2,3], Martin P. Boer[4], Fred A. van Eeuwijk[4], Erik Wijnker[1§], Joost J.B. Keurentjes[1§]

[1] Wageningen University and Research, Laboratory of Genetics, The Netherlands.
[2] Rijk Zwaan, Molecular Biology Research, Fijnaart, The Netherlands.
[3] Managerial Genetics Consulting, Maaseik, Belgium.
[4] Wageningen University and Research, Biometris, The Netherlands.
[†] These authors contributed equally to this work.
[§] Shared last author.

This chapter has been submitted for publication and is online available via BioRxiv: biorxiv.org/content/10.1101/436154v1

ABSTRACT

Chromosome substitution lines (CSLs) are valuable resources to investigate non-allelic genetic interactions. However, the difficulty of generating such lines in most species largely yielded imperfect CSL panels, prohibiting a systematic dissection of epistasis. Here, we present the development and use of a unique and complete panel of CSLs in *Arabidopsis thaliana*, allowing the full factorial analysis of epistatic interactions. A first comparison of reciprocal single chromosome substitutions revealed a dependency of QTL detection on different genetic backgrounds. The subsequent analysis of the complete panel of CSLs enabled the mapping of the genetic interactors and identified multiple two- and three-way interactions for different traits. Some of the detected epistatic effects were as large as any observed main effect, illustrating the impact of epistasis on quantitative trait variation. We have demonstrated the high power of detection and mapping of genome-wide epistasis, confirming the assumed predicted potential of comprehensive CSL sets.

INTRODUCTION

The identification of genetic factors involved in the regulation of quantitative traits is conventionally performed by linkage analysis of genotype-phenotype relationships in segregating mapping populations (*1, 2*). Traditional mapping populations are typically the result of random recombination and segregation of two genotypes in the offspring of an intraspecific cross. Such an approach, however, suffers from a number of inherent complicating factors. These include, amongst others, the simultaneous segregation of multiple quantitative trait loci (QTL) and genetic interactions between them, features that are characteristic for complex polygenic traits (*3*). As a result, conventional mapping populations, such as recombinant inbred lines (RILs), require a large collection of segregating lines to obtain sufficient statistical power to unequivocally detect QTLs and epistasis (*1, 3*). Alternatively, chromosome substitution lines may offer a powerful mapping resource for the systematic dissection of epistatic interactions (*4, 5*).

Chromosome substitution lines (CSLs), a.k.a. consomic strains in non-plant species, differ from established mapping populations by their lack of intra-chromosomal recombination. Consequently, CSLs consist of an assembly of non-recombinant chromosomes, each derived from either one of two genetically different parents (*5, 6*). Genetic variation in CSL populations thus depends exclusively on the reshuffling of complete genotypically distinct chromosomes. As a consequence, the maximum size of chromosome substitution panels, *i.e.* all possible combinations of chromosomes, is finite, depending solely on the chromosome number of the subjected species (2^n, where n is the haploid chromosome number). Complete sets of CSLs offer the advantage of fully balanced allele frequency distributions, providing equal haplotype class sizes in epistatic analyses, and a relatively small population size for species with low chromosome numbers, allowing high line replication in experiments.

To date, a nearly complete set of CSLs has only been established in *Drosophila melanogaster*, due to the ease of generating CSLs and the limited chromosome number in this species (*8*). However, for most other species, complete sets of CSLs are notoriously difficult to generate using conventional backcross approaches and, despite their promises, only a very limited number of CSLs in just a handful of vertebrate and plant species have been developed (*6-10*). Moreover, all these existing panels consist of CSLs with an introgression of only a single donor chromosome in a recurrent genetic background, which considerably restricts the analysis of epistatic interactions. Nonetheless, single chromosome substitution lines (sCSLs) allow the straightforward detection of additive main effects of introgressed chromosomes, while a deviation of the cumulative sum of these effects from the wild type donor phenotype might indicate the presence of epistatic interactions (*5*). However, the

exact strength and genetic architecture of epistasis can only be decomposed by investigating the combined effect of multiple chromosome substitutions.

The recently emerged reverse breeding technology in Arabidopsis determined a major step forward for the development of CSLs (*2, 11*). This approach makes use of the random segregation of non-recombinant chromosomes to the gametes of achiasmatic hybrids, resulting from the transgenic repression of recombination. These gametes are then converted into haploid offspring through crossing to a haploid inducer line (*13*). Finally, the haploid progeny, which consist of an assembly of non-recombinant chromosomes, each derived from either one of the two parents of the initial hybrid, is converted into immortal doubled haploids (DHs). DH seeds occur spontaneously in haploid plants at a low frequency either due to the merging of incidental unreduced gametes that arise by chance, or by somatic doubling. The CSLs produced in this way are now normal diploids containing completely homozygous pairs of chromosomes descending from either parent but in a cytoplasmic background of the haploid inducer line. In Arabidopsis, encompassing five chromosome pairs, a complete biparental panel of all possible CSLs comprises 2^5=32 different genotypes (**Fig. 1**).

RESULTS & DISCUSSION

Here, we report on the construction and application of such a complete set of CSLs resulting from a cross between the Arabidopsis accessions Columbia-0 (Col-0) and Landsberg *erecta* (L*er*). Two of the 32 CSLs resemble the identical genotype of the original parents, albeit both in the cytoplasmic background of the inducer line now (*viz*. Col-0). However, ten CSLs contain a single substituted chromosome (sCSL, 2x5 reciprocally), whereas the other twenty CSLs contain multiple substituted chromosomes (**Fig. 1** and **Fig. S1**). To demonstrate the potential of complete CSL panels in genetic mapping and epistatic analyses, the complete panel was grown in a climate-controlled growth chamber under short day conditions. In order to compare the performance of CSL mapping with conventional linkage analysis a population of RILs derived from the same accessions was grown simultaneously (*14*). All plants were phenotyped for flowering time (days after germination) and main stem length (mm) at the moment of opening of the first flower.

In accordance with the use of conventional consomic strains the additive effect of a single substituted chromosome in comparison to the non-substituted recurrent parental genotype can be analysed. Moreover, since we have generated sCSLs in both recurrent parental backgrounds we can also specifically assess the contribution of epistatic effects to phenotypic variation (*15*). Using a regression model obtained via a backward elimination procedure, significant effects on flowering time were

detected for the substitution of the L*er* chromosomes 2, 3, 4 and 5 in the Col background (**Fig. 2A**; **Table 1**; **Table S1**). Similarly, significant effects on main stem length were observed for the substitution of the L*er* chromosomes 1, 2, 3 and 5 in the Col background (**Fig. 2B**; **Table 1**; **Table S1**). However, in contrast to the reciprocal exchange, the substitution of Col chromosome 3 in a L*er* background displayed no significant effect on flowering time, while the substitution of chromosome 1 did. Likewise, the substitution of the L*er* chromosomes 1 and 3 in a Col background had no significant effect on main stem length, while substitution of these chromosomes in a L*er* background led to significant differences.

In addition to these qualitative background differences, the quantitative effect sizes of the reciprocal substitutions differed substantially. Although the largest effect on main stem length was caused by a substitution of chromosome 2 in both backgrounds, the size of the effect differed approximately four-fold. Furthermore, flowering time was mainly affected by substitution of chromosome 5 in the Col-0 background, whereas the largest effect in the L*er* background was obtained by the substitution of chromosome 2. These differences indicate both qualitative as well as quantitative interaction effects of single chromosome substitutions with the remainder of the genome. Indeed, when the substituted chromosomes and the recurrent background were both included in the regression model, significant interactions of most chromosomes with their background were detected for both traits (**Fig. 2A-B**; **Table 1**).

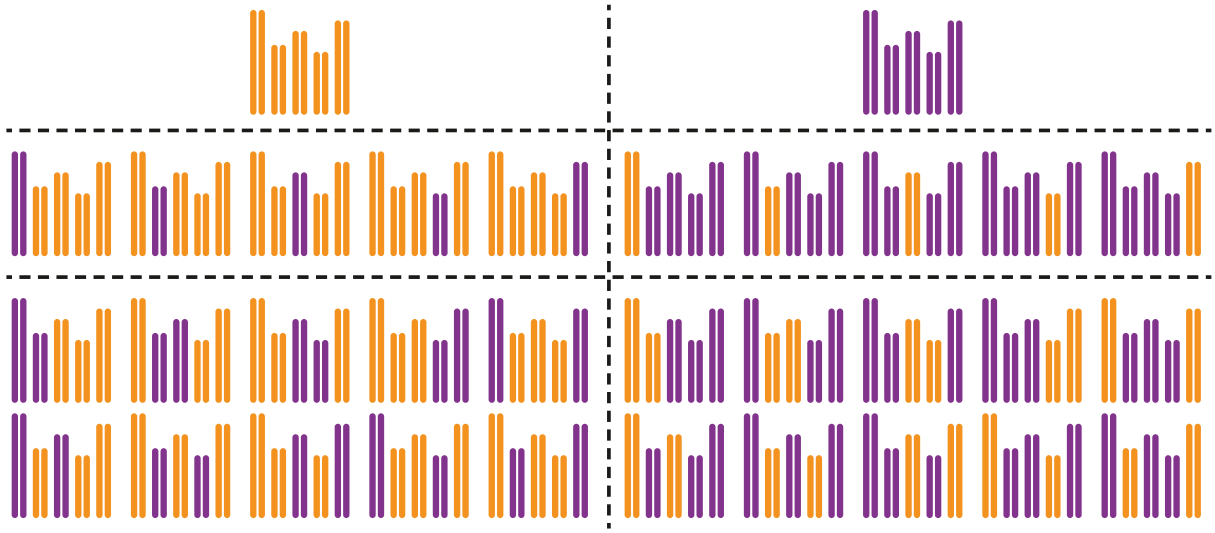


FIGURE 1 | A complete set of *Arabidopsis thaliana* chromosome substitution lines. The complete panel of 32 CSLs can be divided into two reciprocal recurrent backgrounds (vertical dashed line) and subgroups of parental genotypes, single CSLs and CSLs in which two chromosomes are exchanged (horizontal dashed lines). Arabidopsis genomes of each of the CSLs are represented by five homozygous chromosomes derived from either the Col-0 (orange) or L*er* (purple) accession.

Strikingly, the number of QTLs detected in the RIL population using conventional composite interval mapping (CIM) was much lower than in the sCSL panel (**Table S1**), as was also previously observed for rodents (*16*). For variation in flowering time two significant QTLs were detected on chromosome 2 and an additional one on chromosome 1 but no QTLs were detected on any of the other three chromosomes, consistent with previous studies (*17*) (**Fig. 2C**). Furthermore, variation in main stem length in the RIL population is largely explained by a single QTL on chromosome 2, most likely reflecting allelic variation of the *ERECTA* locus (*18*) (**Fig. 2D**).

Despite the high detection power, CSLs inherently offer a low resolution since QTLs can only be mapped to entire chromosomes due to the lack of recombination. To overcome this drawback a reciprocal genome-wide coverage set of near-isogenic lines (NILs) was generated. These were produced by backcrossing sCSLs to one of the recurrent parental accessions and subsequent DH production of recombinant F_1 gametes, as described for the generation of CSLs. In total 413 NILs with either a single or multiple introgressions were generated of which 219 contained a L*er* introgression in a Col background and 194 contained a Col introgression in a L*er* background, as determined by marker-assisted genotyping (**Table S3**). This genetic resource serves to validate and fine-map detected QTLs in the CSLs and confirm possible epistatic interactions with the genetic background.

TABLE 1 | Regression models for different CSL populations explaining variation in flowering time and main stem length. Populations consist of CSLs with only a single substituted chromosome in a particular background plus their recurrent parent, a set of all sCSLs plus recurrent parents, or the complete set of CSLs, including parental genotypes. Regression models contain only backward selected parameters significantly contributing to explained variance. The parameters *Chr1*, *Chr2*, *Chr3*, *Chr4* and *Chr5* denote additive effects of individual chromosomes whereas *BG* denotes background effects. Parameter components separated by a colon indicate interaction effects.

Population	Background	Flowering time	Main stem length
5 sCSLs + Col parent	Col	*Chr2 + Chr3 + Chr4 + Chr5*	*Chr1 + Chr2 + Chr3 + Chr5*
5 sCSLs + L*er* parent	L*er*	*Chr1 + Chr2 + Chr4 + Chr5*	*Chr2 + Chr5*
10 sCSLs + both parents	Col + L*er*	*Chr1 + Chr2 + Chr3 + Chr4 + Chr5 + BG + Chr3:BG + Chr4:BG + Chr5:BG*	*Chr1 + Chr2 + Chr3 + Chr5 + BG + Chr2:BG + Chr3:BG*
32 CSLs	Col + L*er*	*Chr1 + Chr2 + Chr3 + Chr5 + Chr1:Chr3 + Chr1:Chr5 + Chr3:Chr5*	*Chr1 + Chr2 + Chr3 + Chr5 + Chr1:Chr2 + Chr1:Chr5 + Chr2:Chr5 + Chr3:Chr5 + Chr1:Chr2:Chr5*

To demonstrate the complementing value of this NIL population, a subset of reciprocal NILs covering the chromosomes 2 and 5 were grown in similar conditions as the CSLs and RILs. The substitution of chromosome 2 had the largest effect on main stem length, with two-fold longer stems in genotypes carrying a Col chromosome 2 (**Fig. 2F**). Fine-mapping of this chromosome in the reciprocal NILs resulted in a support interval of 7.4 Mbp for the Col set (9.1-16.5 Mbp), while this was much narrower in the L*er* set (9.9-11.3 Mbp). This coincides well with the support interval of the QTL mapped in the RIL population (11.1-11.7 Mbp, **Fig. 2D**) and covers the position of the obvious candidate gene *ERECTA* at 11.2 Mbp. A similar resolution (support intervals of 7.3-8.8 Mbp and 8.0–9.7 Mbp for Col and L*er* NILs, respectively) could be obtained for the fine-mapping of the chromosome 5 QTL for variation in flowering time. No obvious candidate genes are positioned within this support interval although the strong *FLOWERING LOCUS C* (*FLC*) was located on the same chromosome arm (**Fig. 2E**). Surprisingly, despite a ten-day delay in flowering time in sCSLs in which a L*er* chromosome 5 is substituted in a Col background, this QTL was not detected in the RILs (**Fig. 2C**).

An interesting observation from the analysis of the reciprocal NIL sets is the difference in mapping power. The effect on flowering time of a L*er* chromosome 5 substitution in a Col background (ΔFT=-7.4 days) is much larger than *vice versa* (ΔFT=+4.5 days). Likewise, the effect on main stem length of a Col chromosome 2 substitution in a L*er* background is almost eightfold larger than *vice versa*. These differences might reflect discrepancies in effect sizes relative to the recurrent parent's trait value, which might be the result of an accumulation of additive effects, or could indicate a dependency on epistatic interactions. Although the limited set of reciprocal sCSLs also indicates the presence of epistasis, both chromosome 2 and 5 were identified to interact with the background in determining main stem length and flowering time, respectively, the specific origin of these genetic interactions can only be identified by comparing CSLs with multiple substituted chromosomes.

The importance of genetic interactions, relative to the additive effects of single loci, on the phenotypic expression of a trait is part of a long lasting debate (*18-20*) and multiple studies have reported on models including epistasis that explain more variation and have a better predictive power compared to models including only main effects (*22-24*). However, the unbiased testing of epistasis as a source of natural variation is statistically challenging since increasing levels of interaction decrease the number of observations for each genotypic class, which drains the power to detect interacting loci. Furthermore, in most standard mapping populations undetected QTLs and interactions are added to the error term. Finally, overfitting of a model can become a problem due to the close to an infinite number of allelic combinations in a segregating recombinant biparental population. Therefore, most statistical models only include main additive effects and the interactions between them, leaving part of

the heritable variation unexplained (*19*). Completely balanced CSL panels, however, offer the unique opportunity to analyse the relatively limited number of all possible genotypic combinations in a full factorial design and as such provide a more realistic view on the complexity of quantitative trait regulation.

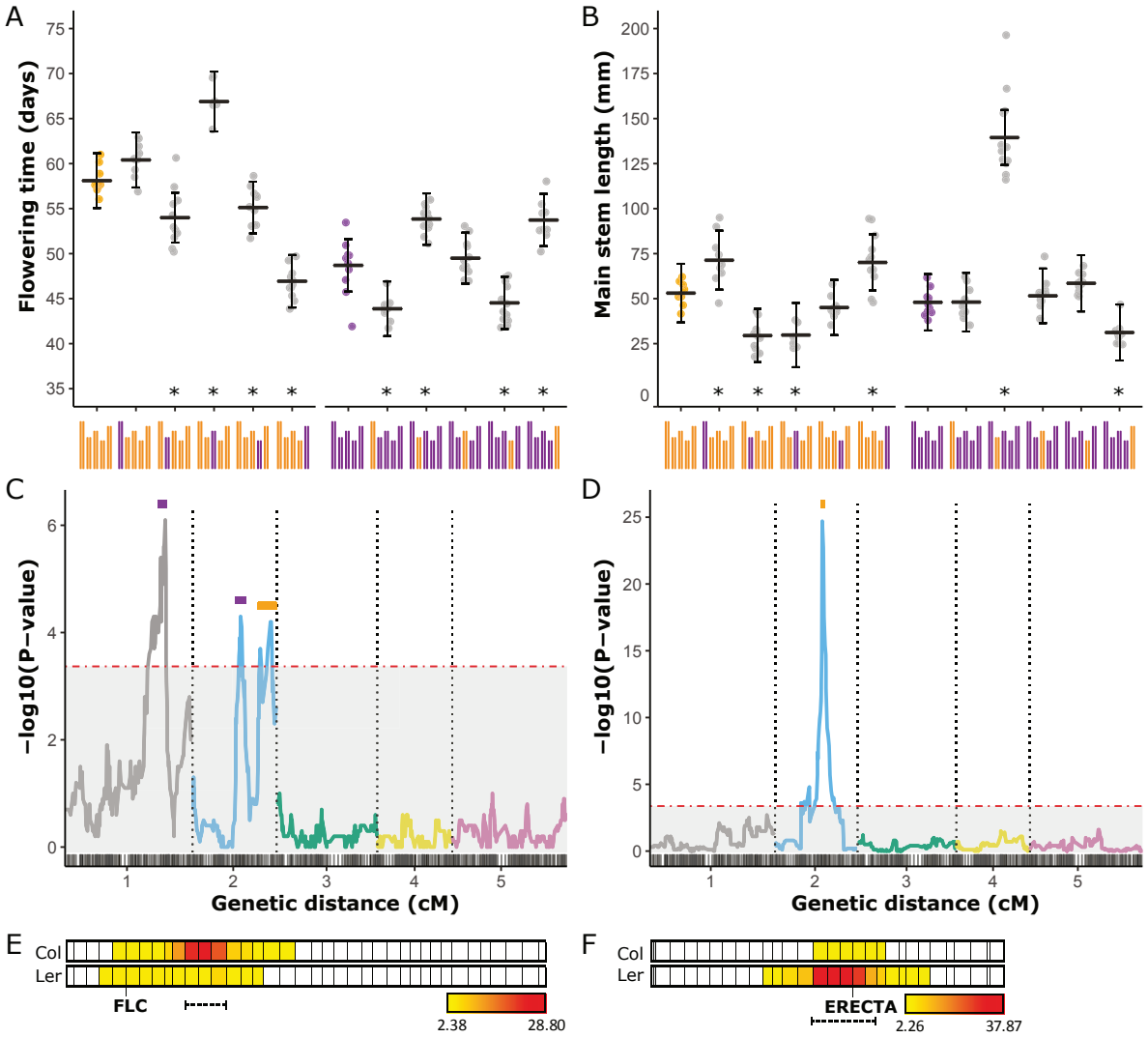


FIGURE 2 | Mapping and validation of single chromosome substitution effects. A-B) Flowering time (A) and main stem length (B) of sCSLs and their recurrent parents. Each dot represents the spatial corrected trait value of an individual of the genotype indicated below the x-axis. Horizontal bars indicate BLUPs with 95% confidence intervals shown as vertical bars (**Table S15**). Asterisks denote significant effects. C-D) QTL plots for variation in flowering time (C) and main stem length (D) as mapped in a RIL population. –log10(P) values for each chromosome are displayed in different colours, while the horizontal red dashed line represents the significance threshold. Support intervals for the QTLs are indicated by coloured bars according to effect sign (orange: $+^{Col\text{-}0}$, and purple: $+^{Ler}$). The x-axis indicates chromosome numbers below a rug profile of the marker positions in cM distance. E-F) Heatmap plots of the effect strength of reciprocal chromosome five introgression NILs on flowering time (E) and chromosome two introgression NILs on main stem length (F). In both panels the upper row represents NIL mapping in a Col background, whereas the lower row represents NIL mapping in a *Ler* background. Vertical lines indicate marker positions in cM. Color intensity from yellow to red specifies the strength of significant effects. Dashed lines below the heatmap indicate support intervals. FLC and ERECTA indicate the position of obvious candidate genes explaining variation in flowering time and main stem length, respectively.

Since clear indications of genetic interactions between chromosomes were obtained from the analysis of reciprocal CSLs and NILs, a regression analysis using a backward elimination strategy on data of the complete CSL panel (**Fig. 3A-B**) was performed to quantify the contribution of epistasis to the phenotype. Using a similar regression approach as was used to test the sCSLs for background interactions, significant chromosome interactions were included in the final model. For variation in flowering time, significant two-way interactions were detected between chromosome 1 and 3, 1 and 5, and 3 and 5 (**Fig. 3C-E**), which partly explain the major effect of genotypic variation of chromosome 5 (**Fig. 3A**). For variation in main stem length a significant three-way interaction between the chromosomes 1, 2 and 5 was detected, while a significant two-way interaction was detected between chromosome 3 and 5 (**Fig. 3F-G**).

Although in general main effect sizes are considered to be larger than interaction effects, here the interaction effect of three chromosomes on main stem length is of similar size as the most effective substitution of a single chromosome (**Fig. 2B**). Most notable for this three-way interaction is a more than 65% increase in main stem length of one genotypic class ($Chr1^{Ler}/Chr2^{Col}/Chr5^{Ler}$) over any of the other seven genotypic classes (**Fig. 3G**). The importance of epistasis is also demonstrated by a comparison of regression models, which either include or exclude epistatic interactions. An inclusive model displays a superior predictive power (R^2 = 0.835) over a model in which epistatic interactions are not considered (R^2 = 0.760; **Fig. S2**). Finally, the impact that genetic interactions can have on the phenotype is illustrated by a case of antagonistic epistasis between chromosome 3 and 5, where the substitution of a Col chromosome 3 with that of L*er* resulted in opposite effects on main stem length, depending on the genotype of chromosome 5.

Our results show that a relatively large part of the observed variation in the analysed quantitative traits can be explained by epistatic interactions. The power to detect these interactions and estimate their effect sizes is greatly enhanced by analysing a complete panel of CSLs, which also includes lines in which multiple chromosomes are substituted. The notion that even for traits dominated by major effect loci (*e.g. ERECTA* in main stem length) epistatic interactions can be revealed, and given the small size of this population, CSL mapping holds great promises for many other quantitative traits in Arabidopsis. There is no reason to assume that similar results cannot be obtained in other species, although larger genome sizes (*i.e.* higher chromosome numbers) might require the simultaneous substitution of two or more chromosomes.

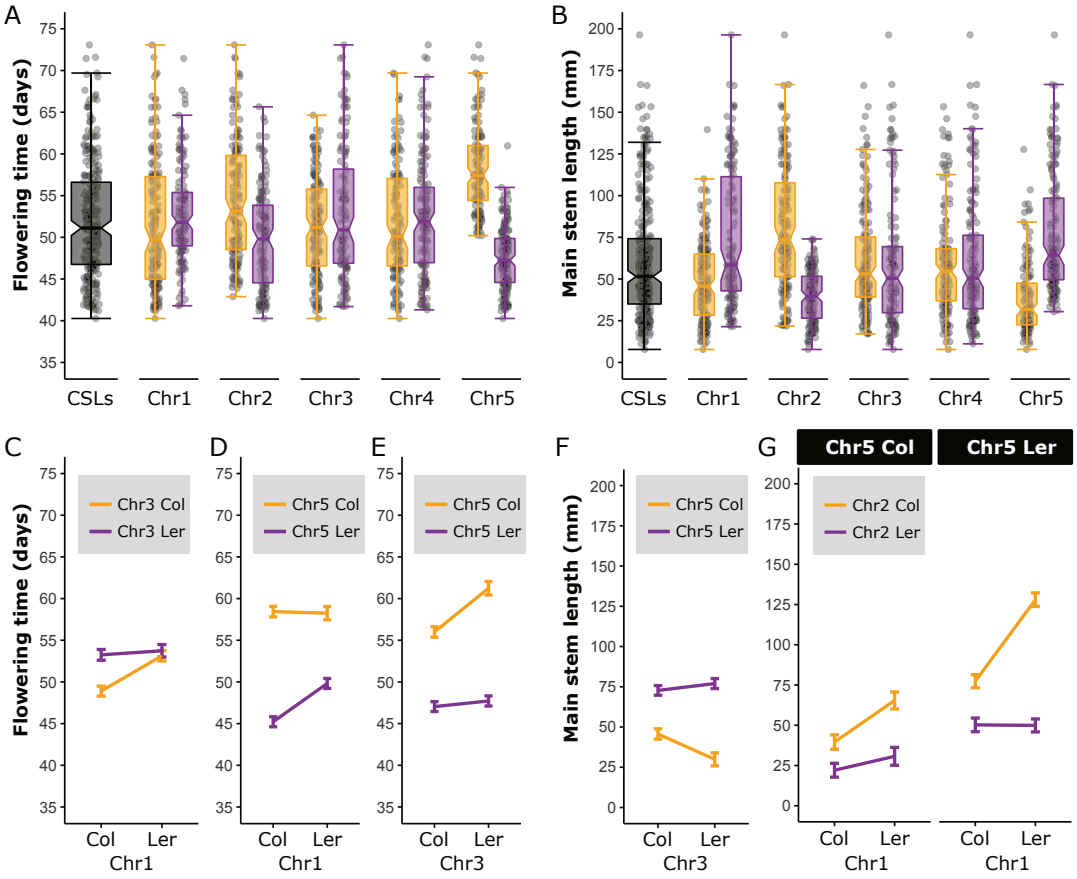


FIGURE 3 | Detection of interchromosomal interaction effects in a complete CSL panel. A-B) Notched box-and-whisker plots of trait values of the complete panel of CSLs for flowering time (A) and main stem length (B). Each dot represents the spatial corrected trait value of an individual plotted in relation to all other individuals (grey boxes) or categorized according to its genotype for the chromosome indicated on the X-axis (orange boxes: Col; purple boxes: L*er*). C-G Regression predicted effect plots of epistatic interactions identified with backward selection models. C-E) Two-way interactions explaining variation in flowering time. F) Two-way interaction explaining variation in main stem length. G) Three-way interaction explaining variation in main stem length. Error bars represent the 95% confidence intervals of the predicted effect.

MATERIAL & METHODS

Development of chromosome substitution lines

Chromosome substitution lines were obtained from crosses between inbred parental lines as previously described (*12*). In brief, semi-sterile Col-0 *RNAi:DMC1* transformed plants, that are impaired in crossover formation, were crossed with wild-type L*er* (CS20) plants to produce achiasmatic F_1 offspring. F_1 plants were then crossed to *GFP-TAILSWAP*, a haploid inducer line, to generate F_1-derived haploids and subsequently doubled haploids (*13*). A number of genotypes that were not obtained by the described

approach were acquired by specific crosses between generated CSLs or between CSLs and parental lines, whether or not containing the *RNAi:DMC1* construct.

Confirmation of genotypes

Potential CSLs were genotyped with a set of 151 SNP markers using KASPar assays (**Tables S3-S4**). These markers covered about 120 Mbp of the total Arabidopsis genome. 95% of the marker intervals were smaller than 2.5 Mbp, which should be sufficient to detect incidental recombinant progeny. To exclude possible phenotypic effects of the *RNAi:DMC1* construct, the absence of the construct in the final selected CSLs was confirmed by additional PCR markers (*25*). During propagation we noted two CSLs ($Chr1^{Ler}/Chr2^{Col}/Chr3^{Ler}/Chr4^{Col}/Chr5^{Col}$ and $Chr1^{Ler}/Chr2^{Ler}/Chr3^{Ler}/Chr4^{Col}/Chr5^{Col}$) exhibiting high intra-line variation. Flow cytometry indicated occasional aneuploidy, suggesting that the plants still carried the RNAi-transgene. Data of these genotypes were excluded from further analyses and the CSLs in the panel were replaced by non-transgenic lines. Removal of these two lines during the data analyses caused non-significant allele frequency distortions of 3.3% at max.

Development of near isogenic lines

Near isogenic lines were acquired by backcrossing the sCSLs to the recurrent parent and by crossing the resulting F_1 to the haploid inducer *GFP-TAILSWAP*. Since the F_1 were transgene-free, it allowed to obtain doubled haploid lines that recombined for a single chromosome. Subsequent genotyping was performed with part of the markers that were available for the confirmation of the CSL genotypes (**Tables S3-S4**). This allowed to fine-map regions with an approximate resolution of 5 cM (~2.5 Mbp).

Propagation

To avoid batch differences introduced by generating CSLs in different series of experiments all lines were first propagated simultaneously in a climate chamber. Seeds were sown on wet filter paper and placed at 4°C in the dark for four days to break residual dormancy and ensure uniform germination. After four days in the cold, plates were transferred to a climate cell at 20°C in the light. After two days, at radicle protrusion, germinating seeds were transferred to 4x4 cm Rockwool blocks in a climate cell set at long day conditions (16h/8h, 20°C/18°C, day/night). Relative humidity was set to 70% and watering was performed automatically with a Hyponex nutrient solution using a flooding system that bottom watered the Rockwool blocks. Five replicates per genotype were sown, and after germination, these were reduced to three well-established replicates. After two weeks of growth, single-leaf samples

were taken for genotyping using three KASP-assays per chromosome (**Table S4**). In addition, a PCR for detecting the presence of the *RNAi:DMC1* construct was performed (*25*). Mature plants were dried and only a single plant was harvested per genotype, which served as the seed stock for the following mapping experiment or any further future experimentation.

Phenotyping experiment

The complete CSL panel was grown in twelve replicates in parallel with three replicates of 100 RILs (**Tables S5-S6**), obtained from the ABRC stock centre (https://abrc.osu.edu/). The handling of the seeds and growth conditions were similar to the propagation conditions, with the exception of short day growing conditions (8h/16h, 20°C/18°C, day/night). Plants were grown in a grid with equal distances between the positions of 12 rows x 60 columns. This grid was divided into three blocks of 12 x 20 each. To create equal number of resources for subsequent analysis, each block contained a single replicate of each RIL (1x100 lines) and four replicates for each CSL (4x32 lines) in a randomized complete block design. In a second separate experiment four replicates of each of the NILs (172 different genotypes in total; **Tables S7-S8**) segregating for chromosome 2 (37 genotypes with Col background and 39 with L*er* background) and chromosome 5 (45 genotypes with Col background and 51 with L*er* background) were grown in the same growth chamber under identical conditions. Here randomized complete blocks consisted of 12 x 30 positions that held two replicates of each genotype.

The number of days after planting at which the first flower opened was recorded (flowering time), at which time point the total length of the main inflorescence was measured (main stem length). Flowering time was corrected for germination date based on daily taken RGB-images by an automated camera system. The day at which the first green leaf could be detected was considered day zero. After three months, the experiment was terminated and plants not flowering by that time were considered outliers due to technical causes and removed from data analysis. Further outliers were determined by image analysis of individual plant growth performance and monitoring reports made during the experiment. Eventually, for most CSL genotypes at least ten replicates were analysed, with a few exceptions of which the CSL consisting of $Chr1^{Ler}/Chr2^{Ler}/Chr3^{Col}/Chr4^{Ler}/Chr5^{Col}$ was most extreme with only four replicates (**Table S9**). For the NILs and RILs only genotypes for which at least two plants were available for each phenotype were included for data analyses (**Tables S9-S10**).

Statistical analyses

The phenotypic data of the RILs and the NILs was corrected for environmental effects using the R packages SpATS (*26*). The script was adapted to our experimental setup, where population and block were included as fixed terms in the model while genotype, row and column were in the random part of the model. The geno.decomp option of SpATS was used to allow for heterogeneous genetic variances for the different populations (respectively the CSLs and RILs in the first experiment and the four different NIL panels in the second experiment). With this model the best linear unbiased predictions (BLUPs) were obtained for the NILs and the RILs (**Tables S11-S12**).

The BLUPs of the NILs and RILs were used as input for the QTL analyses with linear mixed models in Genstat 18th edition. The 676 single feature polymorphism (SFP) markers for the RILs were obtained from previously published data (*27*). Markers with a physical distance of roughly 1 Mbp, corresponding to approximately 5 cM genetic distance in Arabidopsis, were selected (*28*). Genetic predictors between markers were calculated by interval mapping with a step size of 5 cM to bridge any large gaps. For the QTL analyses default settings were used, with minimum cofactor proximity of 50 cM, minimum separation for selected QTLs of 30 cM and Li and Ji threshold settings with genome wide significance levels of $\alpha = 0.05$ (*29*). Initially a single QTL model was fitted. QTLs of the initial analyses were included in the model as cofactors to test for additional QTLs. The QTLs detected and the –log10(p-values) of this composite interval mapping method are reported (**Tables S13-S14**). The support intervals were calculated as a drop of two units in the –log10(p-value) similar to a 2-LOD support interval.

The raw data of the CSLs was corrected for spatial trends with the SpATS R package, and the resulting spatial corrected raw data was used for further analyses (**Table S15**). Individual trait values were preferred over BLUPs for the analyses of the CSLs to increase the degrees of freedom. Either all (for the analyses of the complete CSL set) or a subset (all sCSLs or the sCSLs sharing a single recurrent parent) of the corrected raw data was analysed by applying a backward elimination approach in combination with a multiple linear regression model containing chromosome main effects and two- and three-way epistatic interactions (I).

(I) $$y_{ir} = \mu + \sum_{k=1}^{5} a_k x_{ik} + \sum_{k=1}^{5}\sum_{l>k}^{5} b_{kl} x_{ik} x_{il} + \sum_{k=1}^{5}\sum_{l>k}^{5}\sum_{m>l}^{5} c_{klm} x_{ik} x_{il} x_{im} + \varepsilon_{ir}$$

where y_{ir} is the phenotype of genotype i in replicate r, μ is the overall mean, a_k is the additive effect for chromosome k, x_{ik} is an indicator variable, with $x_{ik} = 0$ $(x_{ik} = 1)$ if chromosome k for genotype i is L*er* (Col), b_{kl} are the effects for the two-way interactions between chromosomes k and l c_{klm}, are the effects of the three-way interactions between chromosomes k, l, and m, and ε_{ir} is the residual error for genotype i in replicate r.

To test three-way epistatic effects, the multiple linear regression model including all main, two- and three-way interactions (I) was compared with a model including main and two-way interactions (II) with backward selection of the AIC criterion using the stepAIC function of the MASS package (with $\alpha = 5.10^{-5}$ to correct for multiple testing) (*30*).

(II) $$h_0 : y_{ir} = \mu + \sum_{k=1}^{5} a_k x_{ik} + \sum_{k=1}^{5}\sum_{l>k}^{5} b_{kl} x_{ik} x_{il} + \varepsilon_{ir}$$

A second step of parameter reduction was used to select the significant two-way interactions for the model with a similar significance threshold. Here, a model resulting from backward selection (IV) was compared to a model including only main effects (III):

(III) $$h_0 : y_{ir} = \mu + \sum_{k=1}^{5} a_k x_{ik} + \varepsilon_{ir}$$

(IV) $$h_1 : y_{ir} = \mu + \sum_{k=1}^{5} a_k x_{ik} + \sum_{k=1}^{5}\sum_{l>k}^{5} b_{kl} x_{ik} x_{il} + \sum_{(k,l,m)\in S_3} c_{klm} x_{ik} x_{il} x_{im} + \varepsilon_{ir}$$

Here S_3 represents the set of the earlier selected significant three-way interactions. Finally, the model including all significant two- and three-way interactions (VI) was tested versus a model consisting of only the mean and the residuals (V):

(V) $$h_0 : y_{ir} = \mu + \varepsilon_{ir}$$

(VI) $$h_1 : y_{ir} = \mu + \sum_{k=1}^{5} a_k x_{ik} + \sum_{(k,l)\in S_2} c_{kl} x_{ik} x_{il} + \sum_{(k,l,m)\in S_3} c_{klm} x_{ik} x_{il} x_{im} + \varepsilon_{ir}$$

Here, S_2 in h_1 represents the significant two-way interaction terms that were selected in the previous round. This backward selection eventually resulted in a model that included all significant three- and two-way interactions and main effects

and all terms underlying the significant interaction terms independent of their own significance according to the principal of marginality.

A similar approach was used for the analyses of the sCSLs were only the main effects model (III) was compared with a model consisting of only the mean and the residuals (V).

For the detection of interactions with the recurrent parental background (either Col or L*er*) all sCSLs were subjected to a similar backward selection procedure. Model (II) was adapted for chromosome x background interactions (VII) and compared with a model for main effects only (V) to test for significant interaction effects between the chromosomes and the background.

(VII) $$h_1 : y_{ir} = \mu + \sum_{k=1}^{5} a_k x_{ik} + b z_i + \sum_{k=1}^{5} c_k x_{ik} z_i + \varepsilon_{ir}$$

Where b is the estimated background effect, z_i is an indicator variable, with $z_i = 0$ $(z_i = 1)$ if the background i is L*er* (Col), c_k are the effects for the interaction between chromosome k and the genetic background. Here significance thresholds were set to $\alpha = 1.10^{-3}$ to correct for multiple testing.

ACKNOWLEDGEMENTS

We like to express our gratitude to F. Becker, G. Stunnenberg, T. Stoker and R. van Genderen of Wageningen University for technical assistance during experimental work. Funding: This work has been financially supported by the Netherlands Organisation for Scientific Research under grant numbers STW-12425 and STW-14389, for which additional support was received from Rijk Zwaan B.V..

REFERENCES

1. C. Bazakos, M. Hanemian, C. Trontin, J. M. Jiménez-Gómez, O. Loudet, New strategies and tools in quantitative genetics: How to go from the phenotype to the genotype. *Annual Review of Plant Biology* 68, 435-455 (2017).
2. C. L. Wijnen, J. J. B. Keurentjes, Genetic resources for quantitative trait analysis: novelty and efficiency in design from an Arabidopsis perspective. *Current Opinion in Plant Biology* 18, 103-109 (2014).
3. J. Bergelson, F. Roux, Towards identifying genes underlying ecologically relevant traits in Arabidopsis thaliana. *Nature Reviews Genetics* 11, 867 (2010).
4. S. H. Spiezio, T. Takada, T. Shiroishi, J. H. Nadeau, Genetic divergence and the genetic architecture of complex traits in chromosome substitution strains of mice. *BMC Genetics* 13, 38 (2012).
5. J. B. Singer *et al.*, Genetic dissection of complex traits with chromosome substitution strains of mice. *Science* 304, 445-448 (2004).
6. J. H. Nadeau, J. B. Singer, A. Matin, E. S. Lander, Analysing complex genetic traits with chromosome substitution strains. *Nature Genetics* 24, 221-225 (2000).
7. M. B. Seiger, The effects of chromosome substitution on male body weight of *Drosophila melanogaster*. *Genetics* 53, 237-248 (1966).
8. A. W. Cowley, M. Liang, R. J. Roman, A. S. Greene, H. J. Jacob, Consomic rat model systems for physiological genomics. *Acta Physiologica Scandinavica* 181, 585-592 (2004).
9. R. Koumproglou *et al.*, STAIRS: a new genetic resource for functional genomic studies of Arabidopsis. *Plant J* 31, 355-364 (2002).
10. J. Kuspira, J. Unrau, Genetic analyses of certain characters in common wheat using whole chromosome substitution lines. *Canadian Journal of Plant Science* 37, 300-326 (1957).
11. E. Wijnker *et al.*, Reverse breeding in Arabidopsis thaliana generates homozygous parental lines from a heterozygous plant. *Nature Genetics* 44, 467-470 (2012).
12. M. Ravi, S. W. L. Chan, Haploid plants produced by centromere-mediated genome elimination. *Nature* 464, 615-618 (2010).
13. C. Lister, C. Dean, Recombinant inbred lines for mapping RFLP and phenotypic markers in Arabidopsis thaliana. *Plant J* 4, 745-750 (1993).
14. C. H. Chandler, S. Chari, I. Dworkin, Does your gene need a background check? How genetic background impacts the analysis of mutations, genes, and evolution. *Trends in Genetics* 29, 358-366 (2013).
15. D. A. Buchner, J. H. Nadeau, Contrasting genetic architectures in different mouse reference populations used for studying complex traits. *Genome Research* 25, 775-791 (2015).
16. M. C. Ungerer, S. S. Halldorsdottir, M. D. Purugganan, T. F. C. Mackay, Genotype-environment interactions at quantitative trait loci affecting inflorescence development in Arabidopsis thaliana. *Genetics* 165, 353-365 (2003).
17. M. C. Ungerer, S. S. Halldorsdottir, J. L. Modliszewski, T. F. C. Mackay, M. D. Purugganan, Quantitative trait loci for inflorescence development in Arabidopsis thaliana. *Genetics* 160, 1133-1151 (2002).
18. O. Carlborg, C. S. Haley, Epistasis: Too often neglected in complex trait studies? *Nature Review Genetics* 5, 618-625 (2004).
19. R. A. Fisher, The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh* 52, 399-433 (1919).
20. R. M. Nelson, M. E. Pettersson, Ö. Carlborg, A century after Fisher: Time for a new paradigm in quantitative genetics. *Trends in Genetics* 29, 669-676 (2013).
21. J. S. Bloom *et al.*, Genetic interactions contribute less than additive effects to quantitative trait variation in yeast. *Nature Communications* 6, 8712 (2015).
22. S. K. G. Forsberg, J. S. Bloom, M. J. Sadhu, L. Kruglyak, O. Carlborg, Accounting for genetic interactions improves modeling of individual quantitative trait phenotypes in yeast. *Nature Genetics* 49, 497-503 (2017).
23. E. Wijnker *et al.*, Hybrid recreation by reverse breeding in Arabidopsis thaliana. *Nature Protocols* 9, 761-772 (2014).
24. M. X. Rodríguez-Álvarez, M. P. Boer, F. A. van Eeuwijk, P. H. C. Eilers, Correcting for spatial heterogeneity in plant breeding experiments with P-splines. *Spatial Statistics* 23, 52-71 (2018).

25. T. Singer *et al.*, A High-Resolution Map of Arabidopsis Recombinant Inbred Lines by Whole-Genome Exon Array Hybridization. *PLOS Genetics* 2, e144 (2006).
26. X. Huang *et al.*, Analysis of natural allelic variation in Arabidopsis using a multiparent recombinant inbred line population. *Proceedings of the National Academy of Sciences* 108, 4488-4493 (2011).
27. J. Li, L. Ji, Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity* 95, 221 (2005).
28. W. N. Venables;, B. D. Ripley, *Modern Applied Statistics with S*. (Springer, New York, NY, ed. 4th, 2002).

SUPPLEMENTARY MATERIALS

Additional supplementary Tables S3-S15 can be found online via: biorxiv.org/content/10.1101/436154v1.supplementary-material.

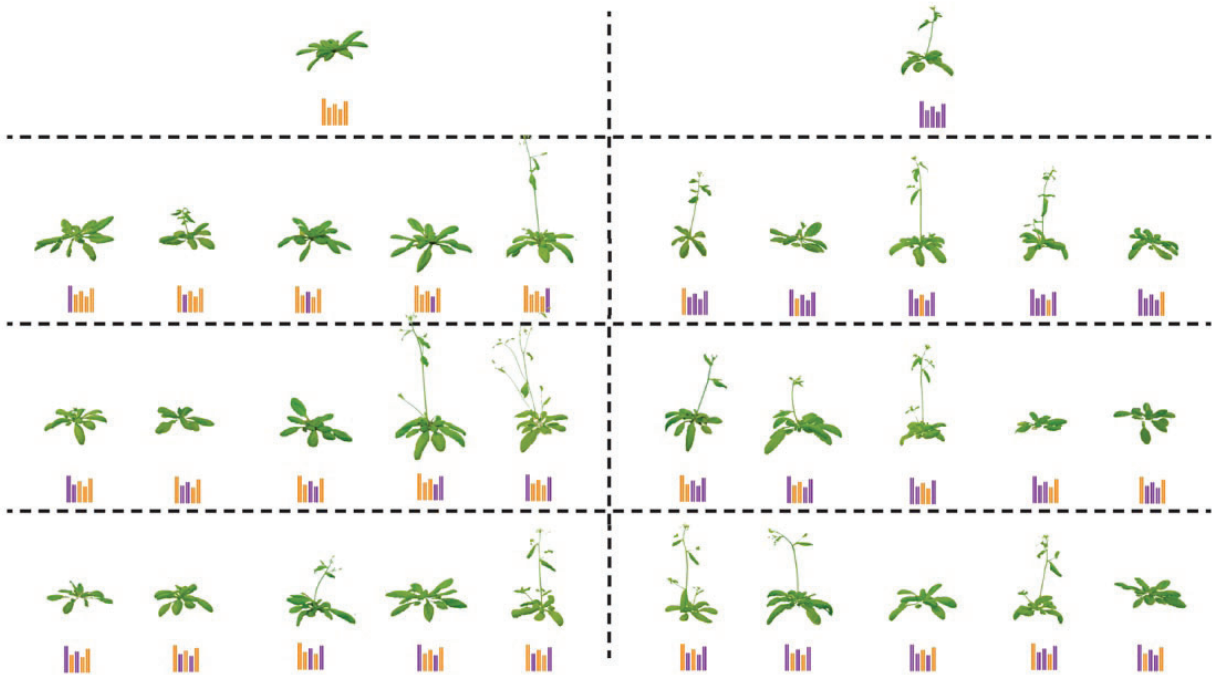


FIGURE S1 | Photographic presentation of phenotypic variation in a complete panel of CSLs. Each image depicts a representative phenotype of the genotype plotted below it. Arabidopsis genomes of each of the 32 CSLs are represented by five homozygous chromosomes derived from either the Col-0 (orange) or L*er* (purple) accession. Depicted plants are of identical age and images were taken at 23 days after sowing in long day (16h light) conditions.

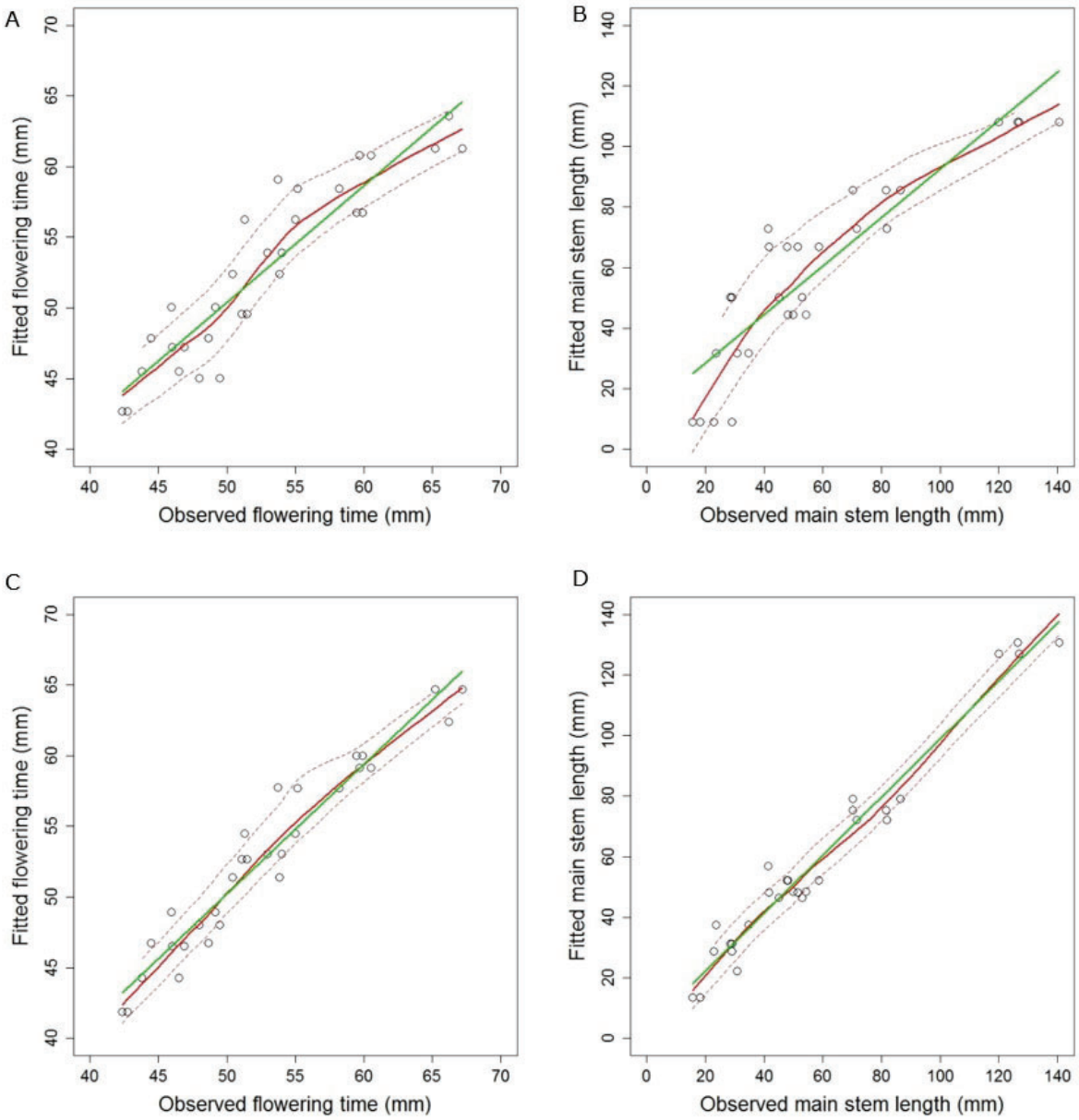


FIGURE S2 | Scatterplots of predicted trait values for models without and with interaction terms. The x-axis shows the observed mean phenotypic values of the CSIs, and the y-axis the predicted values according to the corresponding model. A-B) Prediction of models without interaction terms for flowering time (FT ~ Chr1 + Chr2 + Chr3 + Chr5) and main stem length (MSL ~ Chr1 + Chr2 + Chr5), respectively. C-D) Prediction of the epistatic models for flowering time (FT ~ Chr1 + Chr2 + Chr3 + Chr5 + Chr1:Chr3 + Chr1:Chr5 + Chr3:Chr5) and main stem length (MSL ~ Chr1 + Chr2 + Chr3 + Chr5 + Chr1:Chr2 + Chr1:Chr5 + Chr2:Chr5 + Chr3:Chr5 + Chr1:Chr2:Chr5), respectively. For each plot, the linear regression is shown in green, while the red line represents a trend line of the data including a LOESS-confidence interval between the dashed lines.

TABLE S1 | Detailed overview of main and interaction effects detected in CSL populations. For each significantly detected effect the trait for and population type in which it was detected is given. FT, flowering time; MSL, main stem length; Population type, sCSL: five sCSLs plus their indicated recurrent parent, sCSLs: all 10 sCSLs plus their recurrent parents, All CSLs: all 32 CSLs including the recurrent parental genotypes; Background genotype, the recurrent genotype for the sCSL populations; Chromosome number, the chromosomes for which main or interaction effects were detected, BG: effect of recurrent genotype in sCSL comparisons; Effect size, main effects: effect of background genotype or the substitution of a L*er* chromosome with a Col chromosome (ΔCol-L*er*), interaction effects: the average effect of the substitution of either one of the interacting chromosomes or background compared to the population mean, FT (days), MSL (mm); s.e., standard error of the effect size in the same units; Explained variance, proportion of the population variance explained by each effect; Significance, the significance of the effect.

Trait	Population type	Background genotype	Chromosome number	Effect size	s.e.	Explained variance (%)	Significance (P-value)
FT	sCSL	Col	II	5.4	0.9	2.33	8.7E-08
			III	-7.8	1.1	35.39	1.9E-09
			IV	4.3	0.9	0.05	9.7E-06
			V	12.7	0.9	49.79	< 2E-16
	sCSL	L*er*	I	-5.5	0.8	23.99	1.3E-08
			II	4.8	0.8	18.19	3.1E-08
			IV	-4.6	0.8	24.90	6.7E-08
			V	4.7	0.8	11.77	2.3E-07
	sCSLs		I	-3.6	0.7	1.83	1.3E-06
			II	4.7	0.7	27.66	2.8E-10
			III	-9.2	1.1	0.41	5.5E-14
			IV	2.8	0.9	0.80	2.2E-03
			V	11.2	0.9	41.02	< 2E-16
			BG	6.1	2.1	0.30	5.0E-03
			III:BG	10.3	1.2	11.55	3.2E-13
			IV:BG	6.8	1.1	2.51	7.0E-09
			V:BG	5.9	1.1	2.72	6.8E-07
	All CSLs		I	-1.5	0.6	0.76	9.5E-03
			II	4.7	0.3	10.81	< 2E-16
			III	-7.0	0.6	1.59	< 2E-16
			V	11.2	0.5	62.80	< 2E-16
			I:V	3.8	0.7	1.85	5.8E-08
			I:III	4.8	0.7	3.25	1.2E-11
			III:V	4.6	0.7	2.50	1.1E-10

Trait	Population type	Background genotype	Chromosome number	Effect size	s.e.	Explained variance (%)	Significance (P-value)
MSL	sCSL	Col	I	-23.0	4.1	21.61	8.9E-07
			II	19.4	4.0	19.55	1.1E-05
			III	19.5	5.0	16.55	2.6E-04
			V	-23.0	4.0	16.24	4.5E-07
	sCSL	L*er*	II	88.7	4.0	86.48	< 2E-16
			V	-21.0	4.3	3.68	8.4E-06
	sCSLs		I	-14.5	3.3	10.67	2.5E-05
			II	21.6	4.3	34.27	1.6E-06
			III	21.7	5.4	3.32	1.0E-04
			V	-22.6	3.2	18.15	1.1E-10
			BG	-48.8	7.5	1.88	2.2E-09
			II:BG	63.5	5.6	16.88	< 2E-16
			III:BG	25.5	6.5	1.73	1.5E-04
	All CSLs		I	-25.8	3.5	13.18	2.5E-12
			II	17.5	3.2	33.85	7.3E-08
			III	15.2	2.6	0.00	6.7E-09
			V	-29.1	3.5	24.06	1.6E-15
			I:II	17.2	5.0	8.06	7.1E-04
			I:V	24.8	4.6	0.24	1.7E-07
			II:V	9.6	4.4	2.76	2.8E-02
			III:V	19.3	3.3	1.55	1.4E-08
			I:II:V	33.8	6.6	1.35	4.8E-07

TABLE S2 | Detailed overview of the QTLs detected in RIL and NIL populations. For each significantly detected QTL the trait for and population type in which it was detected is given. FT, flowering time; MSL, main stem length; Background genotype, for the NILs the recurrent background is given, equal allele frequencies are assumed for RILs; Chromosome number, the chromosome on which the QTL was detected; Position, position on the chromosome where the strongest association was detected; Support interval, support intervals were calculated as a drop of two units in the –log10(p-value) surrounding the position of the most significant association; Effect size, effect of the homozygous substitution of a L*er* genotype with a Col genotype at the QTL (ΔCol-L*er*), FT (days), MSL (mm); s.e., standard error of the effect size in the same units; Explained variance, proportion of the total population variance explained by each QTL; Significance, the significance of the strongest association detected.

Trait	Population type	Background genotype	Chromosome number	Position (Mbp)	Support interval (Mbp)	Effect size	s.e.	Explained variance (%)	Significance (-log10(p))
FT	RILs	N.A.	I	23.8	22.2 - 24.1	-2.14	0.64	8.4	6.1
			II	11.2	10.1 - 12.4	-2.68	0.67	13.3	4.3
			II	18.3	15.3 - 19.5	4.18	0.63	32.4	4.2
	NILs	Col	V	8.0	7.3 - 8.8	7.38	0.65	78.2	28.8
	NILs	L*er*	V	8.8	8.0 - 9.7	4.47	0.88	39.7	6.4
MSL	RILs	N.A.	II	11.2	11.1 - 11.7	34.95	3.05	64.0	24.7
	NILs	Col	II	11.3	9.1 - 16.5	8.53	2.30	33.9	3.7
	NILs	L*er*	II	10.6	9.9 - 11.3	65.17	5.02	85.5	37.9

Chapter 4

A genetical-proteomics approach with *Arabidopsis thaliana* chromosome substitution lines provides insights into protein regulation

Cris L. Wijnen[1,2], Twan A.H.P. America[3], Frank F.M. Becker[1], Erik Wijnker[1], Martin P. Boer[2], Fred A. van Eeuwijk[2], Joost J.B. Keurentjes[1]

[1] Laboratory of Genetics, Wageningen University & Research, Droevendaalsesteeg 1, 6708 PB, Wageningen, The Netherlands.
[2] Biometris, Wageningen University & Research, Droevendaalsesteeg 1, 6708 PB, Wageningen, The Netherlands.
[3] Bioscience, Wageningen Plant Research, , Droevendaalsesteeg 1, 6708 PB, Wageningen, The Netherlands.

ABSTRACT

Proteins have long been acknowledged to bridge the gap between the genotype and phenotype. However, contrary to other -omics technologies, like transcriptomics and metabolomics, proteomics has not been extensively used for the detection of genetic effects at the protein level. This is due to the technological difficulties and costs of performing a high-throughput LC-MS/MS shotgun approach. However, the use of a small but genetically diverse panel of Col x L*er Arabidopsis thaliana* chromosome substitution lines (CSLs) might permit the detection of genetic effects at the proteome level. In our experiment, fourteen CSLs were used to enable detection of main and epistatic effects of the various chromosomes. Using strict filtering, the proteomics data revealed more than a thousand associations between quantitative variation in intensity values of 490 proteins and the genotype of specific chromosomes. For an additional 126 proteins the presence of either a Col or L*er* chromosome determined the presence/absence of a protein, indicating possible null-alleles. Genotypic variation at chromosomes 2, 3 and 5 was identified to contribute most to intensity variation of the proteome. For 18% of the proteins interchromosomal effects were detected, accountable for an average of ~25% additional explained variation. The identification of twenty proteins for which intensity variation was completely dependent on interaction terms suggests the presence of reciprocal sign epistasis. The approach described here provides a mean to detect genetic effects on protein intensity, which will help to further understand the complexity of phenotypic trait regulation.

INTRODUCTION

The plant model species *Arabidopsis thaliana* has been extensively explored for genetic and phenotypic variation. These studies range from investigating the effect of induced mutations on gene function in a single accession, to genome wide scans in large sets of natural accessions with high-throughput genotyping platforms. Especially for the accessions Columbia-0 (Col-0) and Landsberg *erecta* (L*er*) a wealth of molecular information is available including *de novo* sequence assemblies (*1, 2*). Thanks to technological advancements and improving data analyses pipelines it has become common practice to use gene expression or metabolomics data as quantitative phenotypes for genetic mapping of genes affecting these phenotypes in genetical-genomics studies (*3-6*).

Surprisingly, there is almost nothing known on differences of protein abundances between genotypes. While variation in a phenotype can be linked to the genotype by using gene expression and protein abundance, gene expression does not always simply correlate with protein abundance. This is ascribed to the large variation in downstream modifications and differences in the rate of generation and degradation of proteins as compared to RNA (*7-9*).

So far, shotgun proteomics for proteome discovery in *Arabidopsis* has been used mainly for detection of specific intensity differences in a single or a few genotypes under contrasting physiological conditions (*10, 11*) or in only a small number of natural accessions (*12*), but has never been used in a systematic mapping study. Over the last two decades, especially sensitivity, resolution and scan speed have drastically improved liquid chromatography-tandem mass spectrometry (LC-MS/MS) based proteomics technology. These improvements now allow shotgun proteomics to be used for genetic mapping due to highly precise measurements of peptide levels and quantitative comparisons between genotypes.

Additionally, a recently developed *Arabidopsis thaliana* panel of chromosome substitution lines (CSLs) is the smallest possible mapping resource in which main and epistatic effects can be estimated (*13*). While many loci segregate simultaneously in each genotype of a standard genetic mapping population, CSLs have a highly simplified genetic architecture due to the substitution of a single or multiple non-crossover chromosome(s) of a donor accession into a reference genetic background (*13-15*). This allows CSLs to be analysed for genetic effects, or quantitative trait chromosomes (QTCs). These genetic effects can be either caused by the introgression of a single chromosome (main effect; QTC_A) or by an interaction between chromosomes (inter-chromosomal effect; QTC_I) (*13*).

In initial studies on the *Arabidopsis* CSLs, analyses were performed on either only genotypes with a single chromosome substitution (these are referred to as single CSLs or sCSLs), or on the complete panel that consists of all CSLs with all possible chromosome combinations of both the donor and recipient parent. For both analyses, multiple QTCs were detected, although interaction effects could only be attributed to the combined effect of specific chromosomes in the analysis of the complete panel (*13*). In mice CSLs, a partial panel, consisting of genotypes with only one or two donor chromosomes, was also shown to allow systematic investigation of interaction QTCs. This allowed the detection of genes that were regulated by epistasis (the interaction between two or more genetic loci) (*16*).

Since epistasis most likely contributes to proteome variation (e.g. proteins are known to act in complexes and as enzymes in linear biosynthesis pathways), here CSLs with either one or two donated L*er* chromosomes in the reference Col genetic background were analysed, in addition to the two CSLs without any substitution (parental-like CSLs; **Table 1**). With such a partial CSL panel at least all the four haplotypes (AA, AB, BA, BB) are included for the estimation of two-way interaction effects. This strategy is similar to the mice CSL panel study and likely to be sufficient to provide an overview of the genetic architecture underlying proteome variation (*17*).

Here we set out to investigate for the first time whether protein intensity data from an LC-MS/MS approach can be used as phenotypes for a mapping study. Essentially, the occurrence of every protein is a phenotype, and thus such an approach yields

TABLE 1 | Overview of the genotypes used in the proteomics experiment. The genotype shows the five-letter code for a genotype (C for Col and L for L*er*). The CSL genotypes that resemble the parents are referred to as parental-like CSL. CSLs with a single or double substitution are sCSLs, or dCSLs respectively.

Genotype	Chromosome					Type
	I	II	III	IV	V	
Col-like						Parental-like CSL
Ler-like						Parental-like CSL
LCCCC						sCSL
CLCCC						sCSL
CCCLC						sCSL
CCCCL						sCSL
LLCCC						dCSL
LCCLC						dCSL
LCCCL						dCSL
CLCLC						dCSL
CLCCL						dCSL
CCLLC						dCSL
CCLCL						dCSL
CCCLL						dCSL

large amounts of phenotypes for which the causal factors of observed variation can potentially be mapped onto chromosomes. The fact that protein regulation is further downstream then gene expression in regulation of morphologic traits, has the potential for proteomics to bridge the genotype-to-phenotype relationship even further. Since genotypically diverse samples are used, in which differential protein intensity could be caused by allelic variation, a differential abundance of proteins (quantitative QTCs) or the presence/absence of a specific protein (qualitative QTC) can be attributed to the substitution of one or more chromosomes.

Plants were grown in a climate chamber and at the moment of flowering the third and fourth rosette leaf from each plant were harvested. The leaves were pooled according to the genotype and subjected to a LC-MS/MS shotgun-proteomic analyses. This resulted in high-quality quantitative intensity data for 1,594 proteins. Of these, 126 proteins showed present/absent variation for some CSLs, while 490 proteins showed quantitative protein intensity variation between the different genotypes that could be attributed to specific chromosome substitutions or combinations thereof. Finally, using an ANOVA-based approach the explained variance by interaction effects was estimated to be approximately 25% of the total phenotypic variance.

MATERIALS AND METHODS

Plant material and growth conditions

Fourteen CSLs of Col-0 x L*er* were grown. These constituted the parental-like CSLs Col and L*er*, as well as twelve genotypes that contained one or two chromosomes originating from L*er* in an otherwise Col genomic background (**Table 1**). Forty-nine seeds were sown for each genotype and seeds were stratified on wet filter paper in Petri dishes at 4°C for 48 hours in darkness. Each seed was sown on a 25x25x40mm GRODAN Rockwool block. These Rockwool blocks were randomized and divided over seven blocks of 12 rows x 9 columns on a the single flooding table. Within each block seven replicates of each of the fourteen genotypes were positioned, leaving ten positions empty.

Protein intensities were measured for three pooled samples for each genotype. Each pooled genotype sample consisted of the left (2 blocks), middle (3 blocks) and right (2 blocks) side of the flooding table. All plants were grown in long day photoperiod (16h light and 20°C starting at 4:00am, 8h darkness and 18°C) with 70% relative humidity and a maximum light irradiance of 200 µmol m^{-2} s^{-1}. The plants were bottom-watered three times per week with Hyponex nutrient solution.

During the growth period, pictures of the top surface of the plants were taken once a day. These were used to monitor the plants and select plants that did not develop properly for removal. When plants flowered (when the first flower had opened), they were harvested daily between 14.00h-16.00h. The third and fourth true leaf of each plant was collected in Eppendorf tubes and quickly frozen in liquid nitrogen. After harvesting samples were stored at -80 °C until further use.

Protein extraction

The leaf tissue was ground by adding 3mm glass-beads to the Eppendorf tubes and shaking them at 300rpm for 2min using a tissue grinder. The Eppendorfs containing the ground tissue were kept in liquid nitrogen and 1ml of chloroform:methanol (C/M; 1:2) was added to each sample. The individual samples of identical genotypes were pooled in 15ml or 50ml tubes per block, creating 14 (genotypes) x 3 (blocks) = 42 pooled samples in total.

Each pooled sample consisted of at least six to a maximum of fifteen replicates per genotype depending on germination success. The pooled samples were vortexed vigorously and immediately thereafter 2ml of homogeneous solution was transferred to a 2ml Eppendorf tube using cut pipet-tips (to allow larger pieces of debris to be taken up). These 2ml samples were sonicated for 10min and placed on an Eppendorf shaker for 10min. Samples were then centrifuged for 10min at max RPM to collect the proteins in a pellet. The supernatant was discarded and the pellet subjected to a second wash step with 1ml of C/M (1:2). These samples were subjected to 10min sonication, 10min shaking on a shaking platform and 10min centrifugation at max RPM. Hereafter, the supernatant was removed as much as possible and a SpeedVac was used to dry the samples completely before storage at -20 °C.

The proteins were extracted from the pellet by adding 0.5 ml of 8 M urea (unless stated otherwise all molar solutions were dissolved in MilliQ-water), supplemented with 50 mM ammonium-bicarbonate (ABC) and 10 mM dithiothreitol (DTT) to stabilize the proteins, and subsequently incubated overnight at 55 °C. Samples were diluted to 6 M urea before the protein concentrations of these stock solutions were measured with a Qubit. Protein concentrations ranged between 1-4 µg/µl per sample. Additional verification of equal protein quantities was confirmed by SDS-PAGE (data not shown).

Protein Digestion

For each pooled protein sample a work solution of 50 µg protein in 50 µl 8 M urea was created. To this, 10 µl of 20 mM iodoacetamide (IAA) supplemented with 50 mM ABC was added and incubated at 37 °C for 30min to prevent disulphide bonding of the proteins. Next, to digest the proteins, 40 µl of LysC/Trypsin (Promega V5073; stock 20 µg dissolved in 200 µl 50 mM HAc, then diluted with 1800 µl 50 mM ABC) was added to the samples, obtaining sample volumes of 100 µl. Both endopeptidases LycC and trypsin cleave proteins at the C-terminal side of specific amino acids (arginine and lysine for trypsin and lysine only for Lys-C). The digest reaction was incubated at 37 °C for two hours.

To extend the trypsin digestion reaction at 1 M Urea, 300 µl of 50 mM ABC supplemented with 0.3 mM CaCl was added to each sample. Trypsin digestion was incubated overnight at 37 °C. Digestion was stopped by adding 45 µl 10% trifluoroacetic acid (TFA) to create a final concentration of 0.1% TFA.

After digestion, peptides were purified using OASIS HLB SPE columns (µElution Plate, Waters, USA) according to the manufacturer protocol. Columns were pre-wetted with 95% acetonitrile, rinsed with 2% acetonitrile (ACN) and 0.1% formic acid (FA) solutions. Next, samples were loaded and subsequently rinsed twice with 2% ACN 0.1% FA. Peptides were eluted with 2 x 40 µl 50% ACN, 0.1% FA and subsequently dried by SpeedVac rotation. Finally, the peptides were dissolved in 40 µl 0.1 M ammonium formate (AF) pH10.

Mass Spectrometry

All forty-two samples were analysed for their quantitative peptide profiles using 2-dimensional nanoLC-HR MS/MS (nanoAcquity-Q-Exactive). The LC-MS/MS system detects peptides in alternating mass spectrometry (MS) mode. This results in high-resolution mass spectra containing information on retention time, mass/charge (*m/z*) value and the relative abundance of each peptide, which allows quantification per sample.

Peptides were loaded in solvent AF onto a first trap column (Xbridge peptide C18 BEH 5 µm particle, 300 µm x 5 mm, Waters, USA) at high pH. Using online two-dimensional nanoLC the peptides were separated into four fractions per sample by increasing ACN concentration (12%, 15%, 24% and 65% in 0.1M AF pH 10). The eluted peptides were on-line diluted 10 times with low pH buffer A (0.1% FA) in order to reduce pH and ACN concentration. The diluted eluate was on-line trapped onto a second trap column (3-µm-bead-packed 2-cm BEH C18 column, Waters USA) prior to

separation on the analytical column using a 60 min gradient of solvent B (100% ACN in 0.1% FA) at a flow rate of 200 nl/min. Analytical separation was performed on a Waters HSS T3 C18 column (1.8 µm particles, 75 µm x 150 mm) on-line connected to a Q-Exactive Plus (Thermo Scientific) mass spectrometer using a nano-electrospray source.

The Q-Exactive instrument was operated in the data dependent mode to automatically switch between full scan MS and MS/MS acquisition. Survey full scan MS spectra (*m/z* 400–1500) were acquired in the Orbitrap with 70,000 resolution (*m/z* 200) after accumulation of ions to a 3×10^6 target value based on predictive AGC from the previous full scan. Precursor ions were selected with charge state 2, 3 or 4 only. Fragmentation of individual peptides in MS/MS mode led to fragmentation spectra containing information on the specific sequence of the peptides.

The eight most intense multiple charged ions ($z \geq 2$) were sequentially isolated and fragmented in the octopole collision cell by higher-energy collisional dissociation (HCD) with a maximum injection time of 60 ms and 17,500 resolution for the fast scanning method, or with a maximum injection time of 110 ms and 35,000 resolution for the sensitive method. Dynamic exclusion was set at 30 s. Typical mass spectrometric conditions were as follows: spray voltage, 2.4 kV; no sheath and auxiliary gas flow; heated capillary temperature, 270 °C; normalized HCD collision energy 28%. The MS/MS ion selection threshold was set to 1×10^5 counts.

Protein identification and quantification

The raw data acquired from the nanoAcquity-Qexactive was used for peptide database searches with MaxQuant v1.5.8.3 software (22, 23). First, the annotation of the peptides was performed by a single protein database search. A combined protein sequence database containing both the Col and L*er* protein sequences (both are sequenced and *denovo* assembled) was provided to MaxQuant, to allow specific protein identification for the different alleles. In MaxQuant, Trypsin and Lys-C specificity was used for database searches, while fixed modification of carbamidomethyl on cysteine and variable modification of methionine oxidation were allowed. False discovery rates were set to 0.01 for PSM, protein and site, and only peptides with a minimal length of seven amino acids and maximum mass of 5,600 Da were accepted.

Second, every peptide identified by MS/MS in at least one of the 42 samples was matched to all other samples by alignment of the MS spectra using the option "*match-between-runs*". This increased the coverage of quantified and identified peptide peaks. For increased specificity the label-free quantification intensity values

(LFQ; or normalized-over-experiment intensity values) of a protein in a single sample was based on the peak intensities of only peptides that were unique to that specific protein. Consequently, LFQ intensities, were thus calculated for each of the 42 samples for only proteins when a unique peptide was identified in at least one sample by MS/MS. These LFQ values are relative values across samples without units that give a measure of the relative proteins abundance. Proteins were reported as "*protein groups*", grouping isoforms or paralogs of the same protein, or in this case, allelic proteins without sequence differences between Col and L*er*, which are all indistinguishable based on peptides sequence.

Selection of high-quality proteins

The LFQ intensities were transformed to log2-values. For each genotype a protein was detected either in zero, one, two or three replicates. When detected in zero replicates the protein was considered to be absent for that genotype (extreme low quantities which do not pass the detection threshold are possible). For each protein the average number of detections per genotype was calculated (excluding genotypes for which the protein was absent). This average number of detections needed to exceed 2.5 for a protein to be included in further data analyses (**Fig. S1**). This reduced the dataset from 3,840 proteins to 1,594 proteins that were detected most consistently over the different samples.

Proteins were considered to be quantitative or qualitative proteins based on the detection of the protein in either all, or a single/few genotypes. Thus the qualitative subset represented present/absent data, where the protein was not detected in all replicates of at least a single genotype. For the quantitative proteins data were imputed for the genotypes for which a protein was detected in only a single or two replicates. The imputation was performed using the rnorm() function in R, where the mean of the observed replicate values of the protein for the specific genotype were taken as the mean of the distribution, and a standard deviation was used that was based on the average standard deviation of the intensity for all 1,468 proteins.

Genetic variance, environmental variance and repeatability (also referred to as heritability) were calculated in R using the repeatability function of the *heritability package* (**Fig. S2**) (*18*). To obtain proteins that showed differential intensity between genotypes, the quantitative proteins were subsequently filtered for a genetic variance > 0.05 as calculated according to the *heritability* package. The resulting 490 proteins were scaled and clustered for a heatmap-plot. The principal components for the PCA plot were constructed based on the protein intensities. The component variables were subsequently tested with an ANOVA for their significant relationship with each of the chromosomes, the two-way interactions of the chromosomes and the replicate

number. The PCA component values were used as the response variables.

The gene ontology analyses were performed with the online tool DAVID where the list of 1,594 proteins was used as a background list (*19, 20*). Distributions of the proteins on the different chromosomes was checked with the online application incorporated in the Thalemine website (*21*).

Chromosome mapping

To obtain the estimated effect sizes of substituted chromosomes and their interactions for each protein, the protein intensity data was subjected to a backward elimination approach based on the AIC criterion similar to the model selection as described in **Wijnen et al 2018** (**Chapter 4**).

The backward elimination approach started from a full model containing all chromosome main effects and two-way interaction effects: $y_{ir} = \mu + \sum_{k=1}^{5} a_k x_{ik} + bz_i + \sum_{k=1}^{5}\sum_{l>k}^{5} b_{kl} x_{ik} x_{il} + \varepsilon_{ir}$. Here y_{ir} is the phenotype of genotype i in replicate r, μ is the overall mean a_k, is the additive effect for chromosome k, x_{ik} is an indicator variable, with $x_{ik} = -1$ $(x_{ik} = 1)$ if chromosome k for genotype i is L*er* (Col), b is the estimated reference background effect, z_i is an indicator variable, with $(z_i = 1)$ $z_i = -1$ if the reference i is L*er* (Col), b_{kl} are the effects for the two-way interactions between chromosomes k and l, and ε_{ir} is the residual error for genotype i in replicate r.

Some modifications to the approach described in **Wijnen et al. 2018** were applied here. First, whereas a complete CSL panel was used before, here the analysis was limited to a partial panel that included the parental-like CSLs and the single and double CSLs in the Col reference background. Therefore the approach was limited to detection of only main effects and two-way interactions, not higher-order interactions.

Second, the partial set of CSLs did not include the sCSL of chromosome three, although two CSLs in which chromosome 3 was substituted together with a second chromosome were analysed. Nonetheless, two-way interactions including chromosome three were excluded because of this genotypic imbalance. This restricts the model to two-way interactions for which enough observations for each genotypic class of the interaction term were present in the current panel.

Third, additional variables were removed from the model based on a significance cut-off of 0.01 for the main effects and 0.005 for the interaction effects, *k* was set to "*qchisq (threshold, 1, lower.tail = FALSE)*" in the stepAIC function of the MASS-

package in R. While k = 2 for genuine AIC, when the threshold is set to 0.01 or 0.005, k = 6.63 or k = 7.88, respectively. Chromosomes and interaction effects that remained in the final model were identified as significantly influencing the abundance of a protein upon substitution of the considered chromosome(s). A significant effect of a single chromosome was defined as a quantitative trait chromosome (QTC) while a significant interaction effect of two chromosomes combined was defined as a QTC_I.

Partitioning genetic variance using regression models

Aside from the above-described final model, two additional regression models were fitted on the same data to partition the total phenotypic variance into different genetic variance components. By including only '*genotype*' as the model factor in a regression analysis, the total genetic variance was estimated ($SS_t = SS_g + SS_e$). The obtained $\%SS_g$ is also referred to as the broad-sense heritability (H^2). The $\%SS_{main}$ or narrow-sense heritability estimates (h^2) were obtained by fitting a model that included only the additive single chromosomes as main effects on the intensity data of each protein. Similar as described in the previous section (**Chromosome mapping**), here the model parameters were selected based on their significance. Using these different models the genetic variation was partitioned into additive, epistatic and unexplained genetic variance ($SS_g = SS_{main} + SS_{two\text{-}way} + SS_{>two\text{-}way}$; **Table 2**). Note that $SS_{>two\text{-}way}$ consists of higher order epistatic interactions and the two-way interactions that were not fitted in the original full model (two-way interactions with chromosome 3).

TABLE 2 | The total phenotypic variance is partitioned into separate genetic components according to the sum of squares of different models. Model 1 estimates the genetic variance ($\%SS_g$) using the total sum of squares (TSS) and the sum of squares explained by the genotype (SS_g). Model 2 estimates only the additive variance (SS_a) and the subsequent $\%SS_a$. In Model 3 the explicit two-way interactions that can be investigated with the current CSL panel are included, obtaining values for the explained epistatic variance (SS_i) and unexplained genetic variance (SS_u) as derived estimates based on the previous models.

Model nr.	Model	TSS formation	Estimated variance component
1	y = genotype + error	TSS = SSg + SSe	SSg
2	y = chr1...5 + error	TSS = SSa + (SSu + SSe)	SSa
3	y = chr1...5 + chr1:chr2...chr4:chr5 + error	TSS = SSa + SSi + (SSu + SSe)	SSa + SSi
	Derived models	**Variance decomposition**	
	Model 3 – Model 2	SSa + SSi - SSa = SSi	SSi
	Model 1 – Model 3	SSg – SSa + SSi = SSu	SSu

RESULTS

MS/MS identifies large proportions of similar peptides between samples and genotypes

We set out to interrogate the different proteomes of fourteen chromosome substitution lines (CSLs) with contrasting genotypes. The fourteen CSLs investigated contained twelve CSLs with either one or two L*er* chromosomes in a Col genetic background and the two parental-like CSLs. All fourteen CSLs were measured in triplicate using biological pooled samples, acquiring a total of 42 samples. Each pooled sample consisted of at least six individual plants that were grown in a randomized block design in a climate controlled growth chamber.

The hydrophobic proteins from each pooled sample were extracted using a chloroform:methanol extraction method (*22*). The subsequent protein-extracts were each digested into peptide mixtures by addition of endopeptidases trypsin and Lys-C. The protein-digests were standardized and purified before injection into the LC-MS/MS system. MS/MS detection is limited to select the more abundant peptides present at any given time during the elution of the sample. Therefore not all eluting peptides are selected for MS/MS fragmentation, however each of the 42 samples was separated into four fractions for higher resolution and increased number of peptide identifications (*23, 24*). The genomes of the CSLs consisted of both Col and L*er* sequences, and to obtain the largest coverage we performed a protein database search including protein sequences of both accessions. First we checked the overlap of identified peptides for all the different samples and genotypes.

In each separate sample fraction ~3,000-4,000 unique peptide sequences were identified by MS/MS, which resulted in ~10,000-13,500 non-redundant peptide sequences for each sample in total. This indicates that the sample fractions consisted of many unique peptides per sample. The identified peptides represented ~13.5% of the total number of submitted MS/MS spectra. The majority of MS/MS spectra did not lead to an identification of a peptide sequence, mainly due to the stringent filters used for signal-to-noise ratios and the presence of peptides that did not fit the search criteria (e.g. different post-translational modifications or non-tryptic/Lys-c digests). Considering all three biological replicates of each genotype together, a cumulative ~21,500 non-redundant peptide sequences were identified per genotype. This points towards a relatively large part of unique peptides per sample. This could either be because different peptides derived from the same proteins were measured, or because of differences in protein profiles between the samples.

The total number of peptides aggregated further into a final dataset of 26,722 unique peptide sequences for all 42 samples together. Of these, 14,912 (57%) peptide sequences were identified in all fourteen genotypes, while only 1,090 peptide sequences were unique to specific genotypes. This indicates a large overlap in identified peptide sequences between genotypes. Although the proteome could be expected to be partly genotype specific at the peptide level due to mutations, it might also reflect that different peptides originate from the same protein or that peptides were below the detection threshold for certain samples. Quantification of the proteome at the protein level is therefore more reliable and informative than at the peptide level.

Obtaining high-quality protein data

The total of 26,722 unique peptide sequences were linked to 4,595 different protein groups present in at least one of the 42 samples. A single protein group consists of one or multiple protein IDs that are indistinguishable based on the identified peptides; these are usually isoforms or paralogs of the same protein, or in this case, allelic proteins without sequence variation between Col and L*er*. Protein IDs were linked to the gene IDs using their Col and L*er* annotations in the database. Protein groups that were linked to multiple different gene IDs were excluded from further analyses to provide a dataset of unambiguously identified proteins. This resulted in 3,840 protein groups with a single linked gene ID.

Based on consistent detection of label free quantification (LFQ) intensities for each protein (ranging from 16.7 to 35.3 and a mean of 25.3), we obtained reliable intensity data for 1,594 high-quality proteins. These proteins were measured in at least a single genotype and linked to a single annotated protein. The distribution of these proteins related to the gene density per chromosome. Given the protein annotation of a single gene, these were distributed according to expected ratios over the nuclear genome (X^2 –p-value: 0.65; **Fig. 1B**). There was, however, an enrichment for chloroplast proteins, which was likely caused by the higher intensities of chloroplast proteins vs nuclear proteins (average intensity values of 25.8 vs 27.9). Overall, the set of 1,594 high quality proteins are a representative selection of the total proteome. Of these 1,594 proteins, 126 were not detected in all genotypes, but in a single or few genotypes. These formed the subset of qualitative proteins. The other 1,468 proteins which were detected in all CSLs were used as the quantitative protein subset.

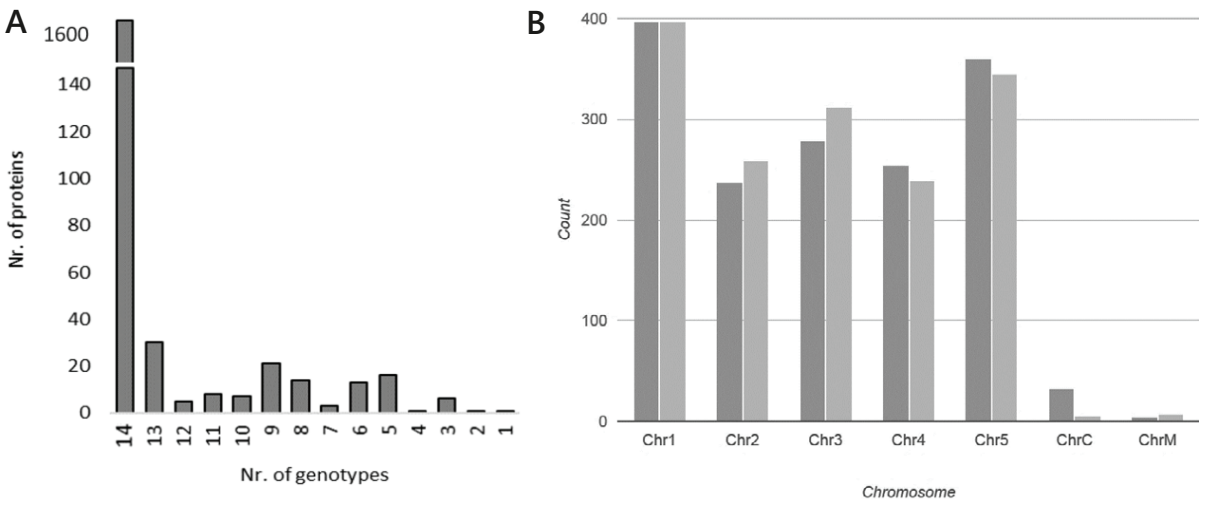


FIGURE 1 | Distribution of the measured proteins over the genotypes. A) The identified proteins follow according to a power-law distribution with 92% of the proteins identified in all genotypes. B) Distribution of protein coding genes over the different chromosomes. Dark grey represents the number of genes on each chromosome encoding detected proteins. Light grey shows the expected number based on the total number of genes on the chromosome and the number of proteins identified.

Chromosome mapping of qualitative proteins

For 93 of the 126 qualitative proteins the abundance pattern over the different genotypes clearly specified which chromosome substitution was responsible for the presence/absence of the protein (**Fig. 2**). According to the gene annotation, the presence/absence of 92 of these could be explained by the genotype of the chromosome harbouring the encoding gene, suggesting cis-regulation. Only variation for a single protein (AT5G20350; TIP1), encoded on chromosome 5, mapped in trans to chromosome 2. While such a trans effect usually is caused by a trans-acting transcription factor, here it might be caused by a translocation of the encoding gene. However, the *de novo* sequence data of L*er* do not provide evidence for this (*2*).

The majority of the 33 qualitative proteins that could not be mapped based on their abundance pattern was either present or absent in only a single genotype. Since the average intensity values of these 33 proteins was much lower compared to the overall average intensities this indicates that these proteins were most probably incorrectly assigned as a qualitative protein due to a failure to detect the protein for a given genotype. However, it could also indicate an extreme case of dependency of presence/absence of the protein on a specific chromosomal combination (i.e. higher-order epistasis).

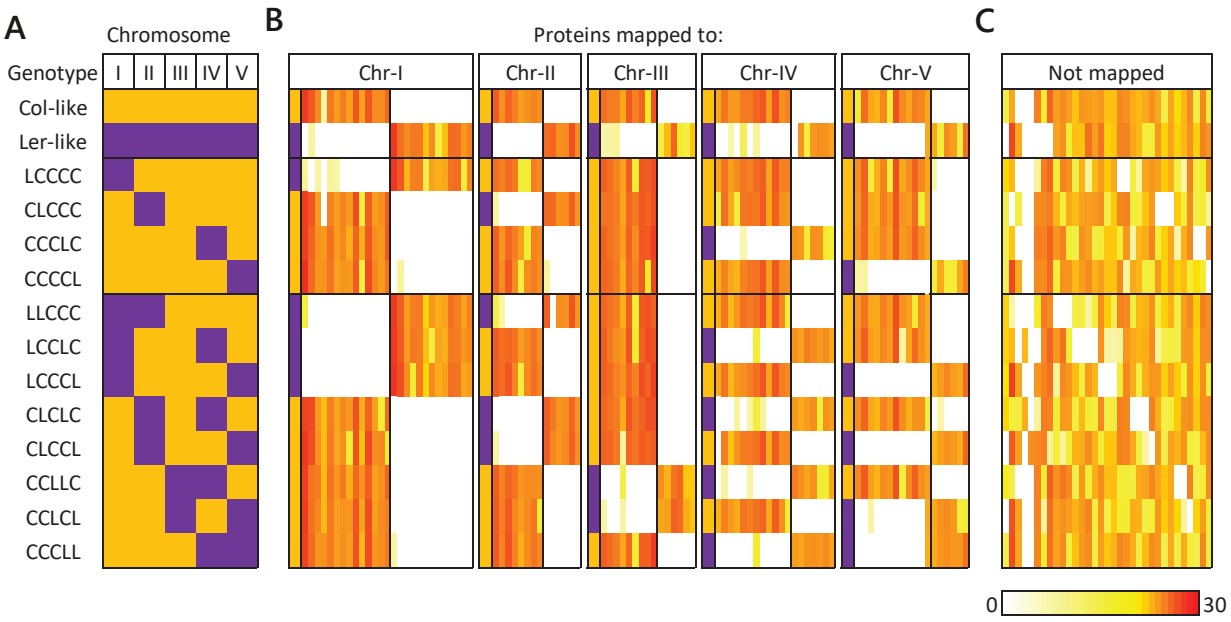


FIGURE 2 | Heatmap of the intensity of qualitatively variable proteins reveals chromosome mapping. A) The chromosomal composition of the different genotypes is shown horizontally and coloured according to the accession of which each chromosome is derived, yellow for Col and purple for Ler. B) The average expression per genotype of each qualitative protein is shown on the vertical axis, coloured according to the expression value. In the first vertical column the genotype at the specified chromosome is shown for each sample to indicate how the expression is linked to the genotype. C) A heatmap showing qualitative proteins that could not be mapped according to their qualitative expression.

Analysis of qualitative proteins

After careful inspection of the qualitative proteins, for twelve encoding genes both a Col and L*er* allele were identified, which were annotated to a different protein sequence and a different identifier in the two original databases. The qualitative protein abundance patterns of these twenty-four proteins showed complementary patterns for the different genotypes (**Fig. 3**). When the intensity values of the proteins encoded by these twelve genes were compared, the abundance of the Col and L*er* proteins was in general significantly different (**Fig. 3C**). Although the occasional non-significant differences do not exclude any functional differences of the proteins, it does indicate that differences in genome sequence do not necessarily have to influence protein abundance.

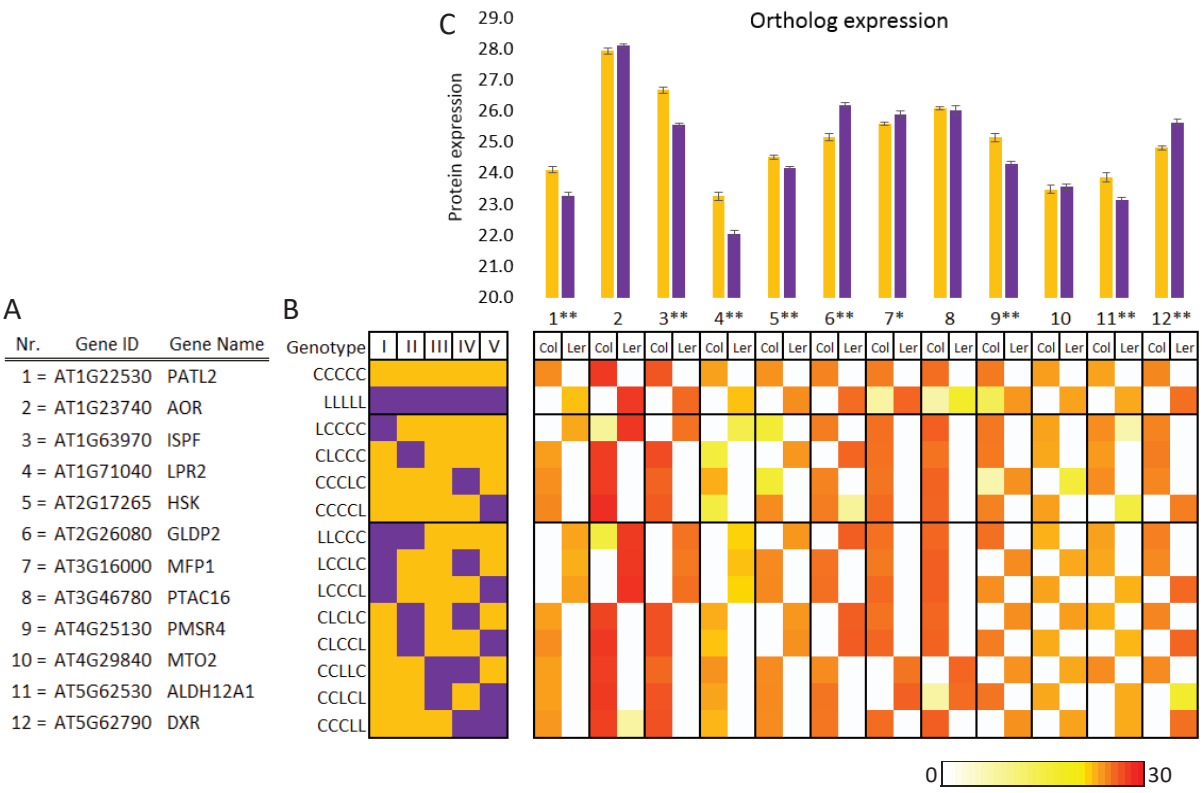


FIGURE 3 | Allelic variants of twelve proteins show significant quantitative variation in their protein abundance. Counter-clockwise from the top-left A) List of genes that contain allelic variation and that have been identified with two separate protein sequences based on the Col and L*er* protein database. The third symbol of the "*Gene ID*" (i.e. ATxGx) contains information on the chromosomal location of the gene. B) On the left-hand side the genotypes are represented horizontally by coloured blocks indicating the parental origin of the chromosome, yellow for Col and purple for L*er*. On the right-hand side a heatmap indicating the abundance of the proteins for each allelic variant of the gene. For instance, the *PATL2* gene is located on chromosome 1 according to the Gene ID (*AT1G22530*) and it is detected as a Col allelic variant only in the genotypes with a Col chromosome 1 or as a L*er* allelic variant in genotypes with a L*er* chromosome 1. C) On top of each heatmap is the average abundance for all genotypes with either the Col (yellow) and L*er* (purple) allelic protein. The number refers to the genes in A and the stars indicate the significance between the Col and L*er* protein abundance tested with a two-sided t-test. One or two stars indicate a p-value <0.05 or <0.001, respectively. The error bars represent the standard error of the mean.

For the remaining qualitative proteins twenty-two were only identified with a L*er* annotation, while thirty-five were only identified with a Col annotation suggesting that sequence diversity between the two parental accessions was causal for the difference in protein abundance patterns. Thirteen of the twenty-two genes with only a L*er* annotation were "hypothetical proteins" for which no Col gene ID was assigned based on sequence similarity. However, none of the qualitative proteins were part of a list of unique genes in the Col or L*er* genome sequences as provided by Zapata et al. (*2*).

Ten of the qualitative proteins were identified with both Col and L*er* annotations, even though the protein was only detected in the presence of either a specific Col or

L*er* chromosome. Most likely these ten proteins were thus identified based only on peptides without sequence diversity between the two accessions, or a mutation was present in the promotor region of the gene or in a transcription factor regulating the expression of the gene.

Several of the encoding genes in the list of qualitative proteins are known to show either expression variation among Arabidopsis accessions or have been identified as the underlying gene of a QTL in a mapping study. A particular example of qualitative proteins is the methylthioalkylmalate synthase (*MAM*) genes located at a single locus on chromosome 5. This locus has first been described as the *GSL-ELONG* QTL (it is related to the elongation of particular glucosinolate backbones) which was detected in the Col x L*er* recombinant inbred lines (RILs) (*25*). Later, this QTL was fine-mapped and confirmed to consist of two genes in Col, the *MAM1* (*AT5G23010*) and *MAM3* (initially named *MAM-like* or *MAM-L*; *AT5G23020*) genes (*26, 27*). The first gene appears as a truncated and non-functional allele in L*er* compared to Col which is confirmed here by the non-detection of the protein in the presence of a L*er* chromosome 5 (*28*). The second, *MAM3*, does not show sequence diversity between Col and L*er* but was identified here as a qualitative protein based on the absence of detection for a single genotype (CLCCC). This is probably caused by a detection threshold issue since there is a clear difference in intensity values between genotypes with chromosome 5 derived from Col (22.5 ± 1.3; to which sCSL-2 belongs) or L*er* (27.2 ± 0.6). This correlates with RNA expression data according to the eFP online browser (*29*), where expression is higher in L*er* compared to Col. *MAM3* is known to produce more elongated glucosinolates, and thus there might be a shift between different glucosinolate products for the two accessions (*30*).

While there was no significant enrichment for genes related to defence response mechanisms, other genes part of the qualitative protein set are known examples of this class. For instance the *GSL-OH* gene (*AT2G25450*), also functional as a glucosinolate modifying enzyme, is known to be functional in the Col genome, while the L*er* sequence contains two deletions that cause the allele to be non-functional (*31*). But also the *PATHOGENESIS RELATED 1* (*PR1* ; *AT2G14610*), *LIPOXYGENASE ISOZYME 2* (*LOX2*; *AT3G45140*), and *ACCELERATED CELL DEATH 2* (*ACD2*; *AT4G37000*) genes show qualitative variation. All are associated with response to pathogens and have been functionally analysed in Arabidopsis (*32-36*).

The qualitative genes show allelic variation, and most of the allelic proteins have significantly different protein abundance. This indicates that these proteins might be in the process of diversification creating orthologous genes. For many proteins this might be a plausible case, because they are related to response to (a)biotic stress.

Genotypic variation results in quantitative intensity differences

Besides the 126 proteins with qualitative intensity differences, 1,468 proteins were identified in all genotypes. After imputation of missing intensity values, these proteins were analysed for their repeatabilities (also known as the broad sense heritability = H^2). Although for many proteins the heritability of intensity was (close to) zero it was moderate to high (>0.5) for a substantial number of proteins, indicating the potential to map protein intensity differences to genotypic variation (**Fig. S2**). However, a high heritability not necessarily reveals a high degree of variation between genotypes. Therefore, the 490 proteins with the largest genetic variance (>0.05) were selected for further exploration and mapping of the chromosome effects.

It was hypothesized that the influence of a single chromosomal substitution might lead to an overal difference in abundance of proteins, and thus initially standard data visualizations were used for exploring the complete data set. In a heatmap, proteins did not cluster according to the genotype of specific chromosomes (**Fig. 4**), but patterns of protein abundance indicated a combinatorial effect of chromosomes 2 and 5 on many proteins. Note for instance how there seems to be a correlation between genotypes with chromosomes 2 and 5 Col and a block of higher abundant proteins on the left of the heatmap, and lower abundant proteins on the right of the heatmap (most notable in samples LCCCC, CCCLC, LCCLC and CCLLC).

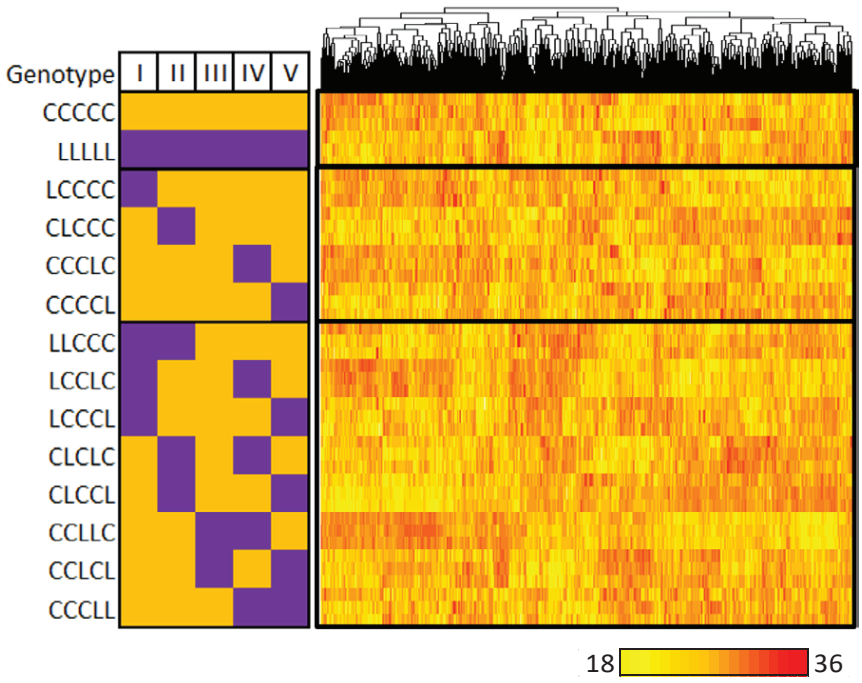


FIGURE 4 | The genotypes are distinguishable based on their protein abundance profiles. On the left, the different genotypes used are shown. Each genotype is sampled three times. On the right, the scaled intensity values of 490 proteins (in vertical direction) were clustered according to their protein intensity for each sample. The three replicates of the genotypes show differences, but have in general a similar pattern.

A principal component analysis (PCA) of the 490 proteins also indicated that the first components were mostly affected by the genotype (**Fig. 5**). The first and second component explain 26.6% and 12.6% of the total variance, respectively, and only the first five components account for more than 5% (total 59.7%). Especially, chromosomes 2 and 5 contribute to separation at the first component (both with a p-value < 10^{-13}) confirming the clustering of proteins in the heatmap (**Fig. 4**). Chromosome 3 contributes most significantly (p-value < 10^{-15}) to the second dimension in the PCA. Considering the explained variances of the PCA components, this indicates that especially chromosomes 2 and 5 together provide more variation than the other chromosomes do. Specifically in the combination $Chr2^{Col}$-$Chr5^{Col}$, much more within group variation seems to be present than is the case for the other three haplotypes between chromosome 2 and 5.

The large contribution to variation in protein abundance of the substitution of chromosomes 2 and 5 can be related to the observation of QTL hotspots on chromosomes 2 and 5 in a previous study (*37*), although this was performed in a L*er* x CVI RIL population. In this study an enrichment of QTLs for proteins, metabolites and phenotypes on chromosome 5 was found while a strong QTL hot spot for RNA expression, metabolites and phenotypes was present on chromosome 2, located in close proximity to the *ERECTA* gene (*AT2G26330*) (*38*). The *ERECTA* gene is pleiotropic for many traits and thus could potentially explain many of the QTLs (*39,*

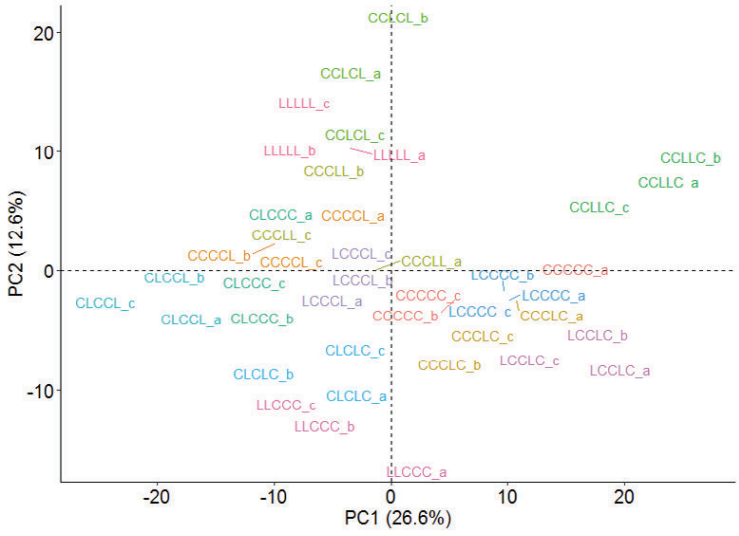


FIGURE 5 | PCA analysis indicating a strong genetic correlation for the measured samples. The three replicates per genotype have the same colour, and the suffix "a", "b" and "c" indicate the three replicate numbers. On the left side of the PCA plot genotypes with chromosome 2 L*er* and/or chromosome 5 L*er* are positioned, while on the right side all genotypes with both chromosomes 2 and 5 Col are clustered. Note also how the genotypes contributing most to the first two components were those genotypes that had two chromosomes substituted, indicating that variation in protein abundance increases with more genetic substitutions.

40). Although there might be a similar biological explanation for the variation caused by chromosome 3, it is more likely that the contribution of chromosome 3 variation is also caused by the unequal allele frequencies in the population (in the current set only three genotypes contained a L*er* chromosome 3).

Mapping of main effect QTCs for quantitative protein abundance

With the indication that chromosome 2 and 5 contribute most to the total variation of the proteome, it was considered that this should also be reflected at the individual protein abundance levels for these 490 proteins. Using an ANOVA-based approach (**see material and methods**), we checked what the effect was of a substitution of one (main effect) or two (interaction effect) chromosomes on the protein abundance.

With the present CSL set we estimated only the main effect of each chromosome and six interaction effects: the interactions with chromosome 3 were excluded. A "background" term was included that should capture most of the genetic variation caused by the four two-way interactions with chromosome 3 and higher-order interactions. For most proteins a significant effect was attributed to the substitution of one of the chromosomes (referred to as QTC_A), or an inter-chromosomal interaction (QTC_I).

For the 490 proteins a grand total of 1,266 QTCs were detected. While for 53 proteins no significant QTC was identified, for 110 proteins a single QTC was detected that linked the protein abundance directly with the substitution of the specified chromosome. For only 49 of these proteins the gene was located on the substituted chromosome itself, indicating a local effect of the substitution, whereas in 61 cases the observed variation was due to a substitution in trans.

The two proteins with the highest significant effect (p-value < 10^{-15}) were encoded by the genes *SEDOHEPTULOSE-BISPHOSPHATASE* (*SPBase*; *AT3G55800*) and *EPITHIOSPECIFIER PROTEIN* (*ESP*; *AT1G54040*). For both, protein intensities were affected by substitution of the chromosomes harbouring their encoding genes, indicating that the variation in protein abundance was caused by different alleles or promotor sequences of these genes. *ESP* was originally mapped as the *TASTY* locus in the Col x L*er* RIL population, conferring resistance to insect feeding in L*er* (*41*). The *ESP* gene was later cloned and confirmed to be responsible for hydrolysis of specific glucosinolates (*42*). The *SPBase* gene is described as an enzyme with a role in photosynthetic carbon fixation where it influences multiple metabolic and developmental processes (*43*). Although not much is known about natural variation in expression of this gene, according to the 1001 genome browser there is extensive allelic variation present for the Col and L*er* sequences which could lead to protein abundance differences (*44*).

For most proteins it was not exceptional to detect multiple QTCs. This large set of proteins contained for instance proteins with multiple QTC_A, proteins with a single significant interaction QTC_I (i.e. for every QTC_I the two main effect chromosomes of the interaction were included as QTC_A), but also 45 proteins for which more than five significant QTCs were detected (**Fig. 6A**). Of the grand total of 1,266 QTCs, there were 1,084 QTC_A (mean effect size: 0.49 fold difference ± 0.44 s.d.) and 142 QTC_I (mean effect size 0.83 fold difference ± 0.39 s.d.), which reflects that QTC_I need to be of considerable effect size to be detected compared to main effects.

QTCs were detected for every possible chromosome combination, but an above-average number of QTC_A were detected for chromosomes 2 and 5 (**Fig. 6B**). This is consistent with the observation of a highly significant influence of chromosomes 2 and 5 on the total proteome (**Fig. 4 & 5**) and thus this indicates again that especially the substitution of these two chromosomes is important for variation in protein intensities.

Chromosome substitutions can have large effects on the abundance of proteins

For thirty proteins the substitution of a chromosome led to at least a one fold-change in abundance according to the ANOVA-based estimated effect sizes (**Fig. 7**). Similar to most of the genes associated with qualitative abundance variation, many of the genes encoding the proteins with fold-change abundance differences have been suggested to influence phenotypic variation, especially in response to (a)biotic stimuli.

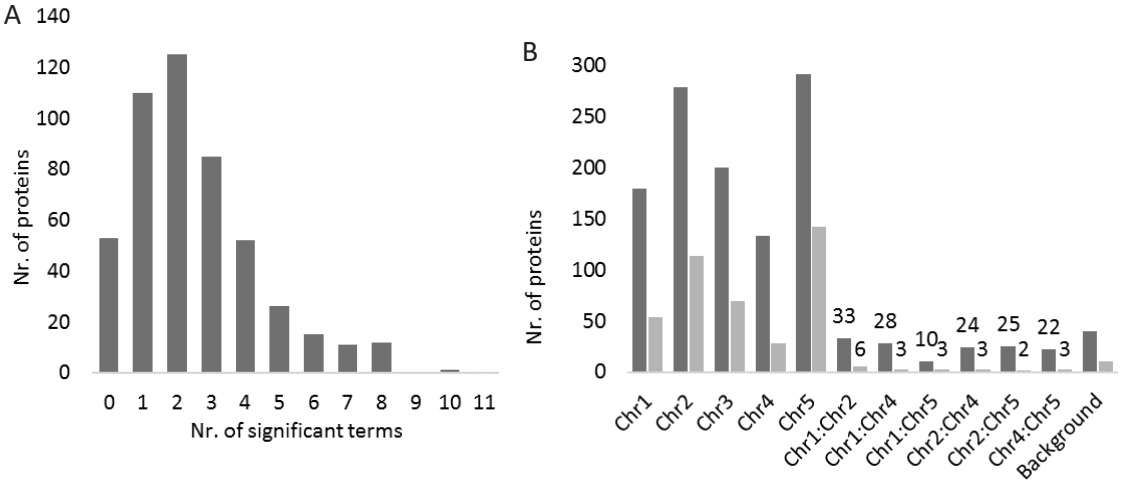


FIGURE 6 | Distributions of the significant protein abundance QTCs. A) The number of significant effects identified per protein in the final model of the protein. For most proteins only one or two significant chromosomes were identified, while there were also many proteins that had multiple significant effects, including two-way interactions between chromosomes. B) Distribution of the number of identified effects per individual term in the final ANOVA models. Dark-grey shows the number of significant effects detected according to the specified term in the final model. Light-grey shows how often the specified term was the most significant term in the final model.

The largest effects (>4 fold-change) were observed for the *ESP* (*AT1G54040*), *SPBase* (*AT3G55800*), *LIGHT HARVESTING COMPLEX PHOTOSYSTEM II* (*LHCB4.2*; *AT3G08940*) and *EPITHIOSPECIFIER MODIFIER 1* (*ESM1*; *AT3G14210*) proteins. *ESP* and *ESM1* are known to have counteracting roles in the hydrolysis of glucosinolates that allow fine-tuning of the production of indol-3-acetonitrile, a metabolic compound that deters most generalist insects *(42, 45, 46)*. Although these two proteins together regulate the abundance of indol-3-acetonitrile in a non-additive fashion, they are themselves predominantly regulated by a main effect.

Additional genes that respond to (a)biotic stimuli included *MYROSINASE-BINDING PROTEIN 2* (*MBP2*; *AT1G52030*; glucosinolate degradation), *PATHOGENESIS-RELATED 2, 4* and *5* (*BGL2*; *PR4* and *PR5; AT3G57260, AT3G04720* and *AT1G75040*; pathogen response), *COLD-REGULATED 15A* and *47* (*COR15A* and *COR47*; *AT2G42540* and *AT1G20440*; cold response), *EARLY RESPONSIVE TO DEHYDRATION 10* (*ERD10*; *AT1G20450*; drought response) and *PLANT CADMIUM RESISTANCE 1* (*PCR1*; *AT1G14880*; cadmium response). Also here, similar to the qualitative proteins, the largest contribution to proteome

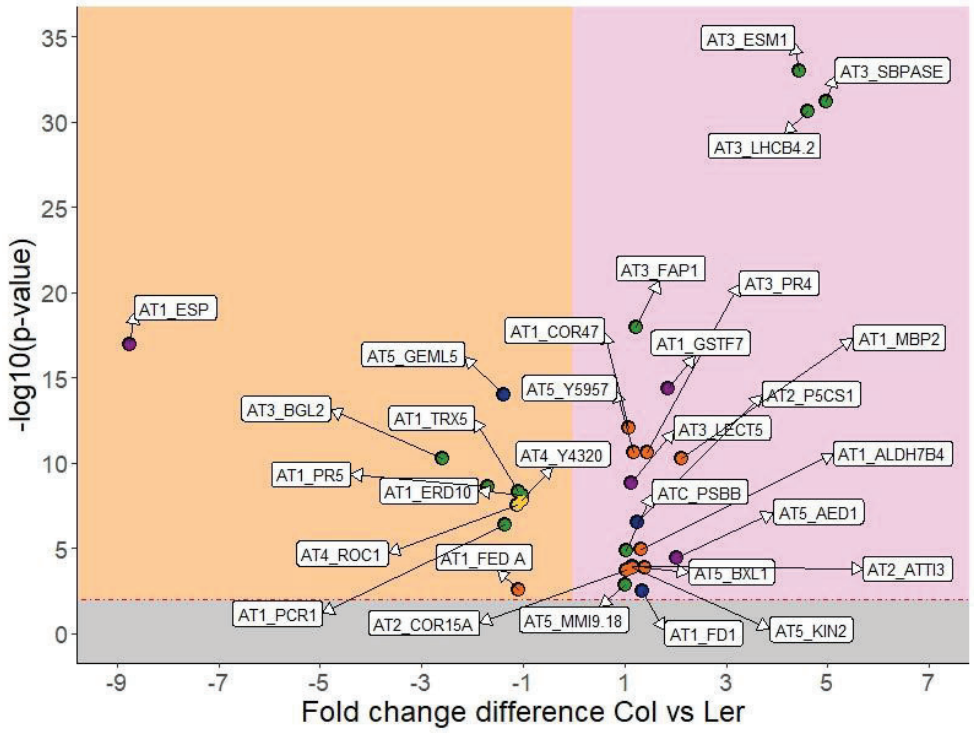


FIGURE 7 | The genes with the highest fold-change protein expression differences. Fold-change differences are represented on the x-axis, where a negative value indicates the protein is higher expressed in the genotypes with a Col chromosome (also indicated by the orange background colour), while a positive value indicates the protein was higher expressed in the genotypes with a L*er* chromosome (purple background). The y-axis indicates the significance for the ANOVA F-test. The red-dashed line indicates the significance threshold at p-value = 0.01. The different colours of the proteins indicate which chromosomal substitution led to the differential expression of the protein, purple, orange, green, yellow, blue for chromosomes 1-5, respectively. Each protein is labelled according to its gene name annotation, where the first three characters indicate the chromosome.

variation in the two parental accessions is by proteins and enzymes that are either known to produce a diverse response to a specific stress or involved with steps at the end of metabolic pathways. This again advocates the hypothesis that especially such genes are likely to contain mutations that can create variation within species.

According to the gene annotation of each protein, a QTC_A affected either the protein abundance by their encoding genes located on the substituted chromosome (local QTC, predominantly cis-regulation) or by genetic factors located on another chromosome (distant QTC, trans regulation). The number of QTC_A did not always reflect the number of genes on the different chromosomes: While for each chromosome there was enrichment of local versus distant QTCs, the substitution of chromosome 1 or 4 led to significant higher than expected local QTCs compared to distant QTCs (both with X^2 –p-values <0.001; **Fig. 8**). The finding that there are more local QTCs than distant is not unexpected, since the local QTCs contain both trans and cis effecting genes, while distant QTCs only contain trans effects. Besides, it is more common that cis QTLs have larger effects than trans QTLs. Therefore chromosomes 1 and 4 might actually follow a more expected distribution than chromosomes 2, 3 and 5, which have more distant QTCs. This shows that substitution of chromosomes 2 and 5, and here also chromosomes 3, affect the protein intensities of genes located not only in cis, but especially in trans. This likely indicates that variation in important trans regulating genes on these chromosomes are affecting the proteome differences.

Description of the interaction effect QTCs

The 142 QTC_I detected with the ANOVA-based approach were distributed over 89 proteins (18%; **Fig. 6B**). An additional forty background effects were detected, indicating the presence of either interactions with chromosome 3 or higher-order interactions. Although the interaction between chromosomes 1 and 5 was only significant for ten proteins, for the other interactions at least twenty-two significant observations were made.

The number of QTC_A at each contributing chromosome was not indicative for the number of QTC_I between those chromosomes. For instance, the significant interaction between chromosomes 2 and 5 was not more frequently observed than significance of other interactions. There were 199 proteins with both a chromosome 2 and 5 main effect, while only 25 of these had a significant QTC_I for chromosome 2 and 5.

When instead of all detected QTCs for each protein only the most significant term of the ANOVA-model was considered, a similar distribution for the QTCs was observed over the chromosomes (**Fig. 6B**). Still twenty proteins were identified for which the interaction between chromosomes was the most significant, indicating mainly an epistatic regulation

of the quantitative expression of these proteins. The regulation of protein variation thus seems to be mainly affected by the substitution of single chromosomes, although for several inter-chromosomal interactions a strong effect can be observed.

The largest interaction effect was observed for *PATHOGENESIS-RELATED PROTEIN 2* (*PR2/BGL2*; *AT3G57260*) where the combination of chromosomes 4^{Ler} and 5^{Ler} led to a 2.5-fold increase in PR2 protein abundance. In addition, PR2 was also strongly regulated by a local substitution effect of chromosome 3^{Ler} (effect size 2.6-fold increase), indicating that regulation of this protein might be far from straightforward and depending on multiple factors. For a high PR2 protein abundance at least chromosome 3^{Ler} needs to be present, and subsequently the combination of chromosomes 4^{Ler} and 5^{Ler} lead to an even higher expression. Still, the L*er*-like CSL does not display the highest abundance of this protein, indicating that an additional effect of another chromosome (combination) decreases the total effect. Another example is the RNA-binding protein *ALWAYS EARLY 3* (*ALY3*; *AT1G66260*). Specifically the combination of chromosomes 1^{Ler} and 2^{Ler} leads to a two-fold reduction of the ALY3 protein quantity, and here no other interactions or main effects interfere. ALY3 modulates plant growth and development processes by transporting mRNA from the nucleus to the cytosol as part of the transcription-export-complex (TREX). This TREX-complex consists of multiple proteins that could also epistatically regulate the abundance of *ALY3*.

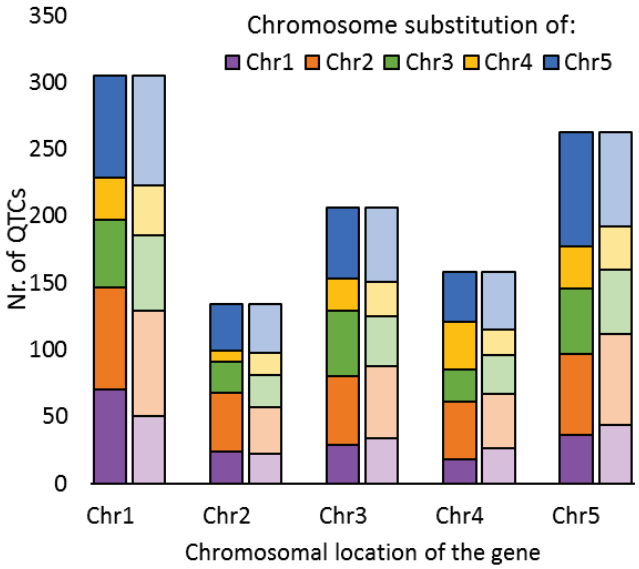


FIGURE 8 | QTCs according the location of the genes affected and the specific chromosomal substitution. Each single dark-coloured stacked bar shows the total count protein-coding-genes on the specified chromosomes that were effected by a QTC (i.e. a total of 305 protein-coding-genes located on chromosome 1 had a significantly detected QTC). The different colours indicate the chromosome substitution that was causal for the QTC (i.e. only a small proportion of the protein-coding-genes located on chromosome 1 were affected by a substitution of chromosome 4). The transparent stacked bars indicate the number of expected QTCs according to a χ^2-distribution table. Chromosomes 1 and 4 show a significant deviation from this expected number of QTCs (χ^2 -p-values: 0.016; 0.479; 0.293; 0.001; 0.333 for chromosomes 1-5 respectively).

Epistasis can explain a large part of the genetic variance

For each of the 490 proteins three different genetic models were considered (**Table 2**). These different models allowed the partitioning of the genetic variance ($\%SS_g$) into different components, the main effects (or additive variance; $\%SS_a$), the two-way effects (or epistatic variance; $\%SS_i$) and the higher-order effects (or unexplained genetic variance; $\%SS_u$). Note again that we only estimate six two-way interactions instead of all ten possibilities, and thus the four two-way interactions including chromosome 3 are included in $\%SS_u$.

Using this approach the main effects accounted for 49% (± 19.5% s.d.) of the total variance on average for proteins where significant main effects were identified (**Fig. 9**). Similarly for the 89 proteins including an interaction term, two-way epistasis accounted for 24.7% (± 16.4% s.d.) of the total phenotypic variance on average (or 35.8% ± 25.4% s.d. of the genetic variance = $\%SS_i/\%SS_g$). For nineteen proteins part of the genetic variance was only explained when interaction effects were included in the model (**Table 2**; model 3). Indeed, for these proteins no main effect was detected, but the combination of two chromosomes led to a significant difference in protein abundance between the genotypes. This type of model is associated with sign epistasis (i.e. a crossover interaction) where the opposite allelic (or chromosomes) combination leads to a similar effect.

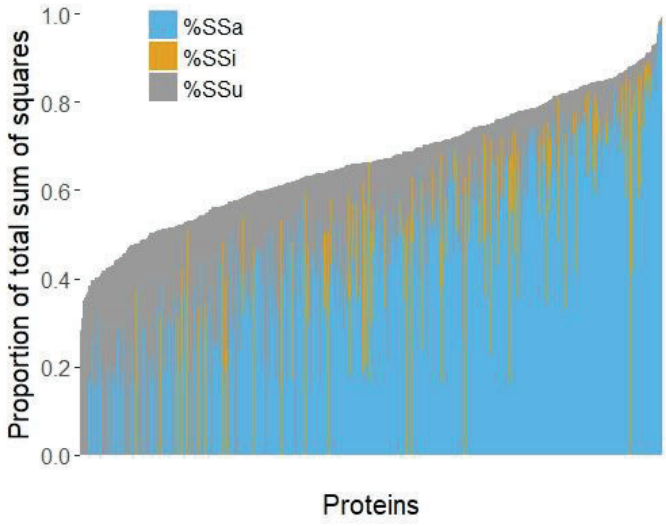


FIGURE 9 | Both main and interaction effects can explain a large part of the genetic variance. Every vertical line represents the total variance of an investigated protein; here a total of 490 proteins were subjected to the analyses. For each protein the genetic variance ($\%SS_g$) is partitioned into the additive variance ($\%SS_a$; blue), two-way epistatic variance ($\%SS_i$; orange) and the remaining unexplained genetic variance ($\%SS_u$; grey). The environmental, or residual variance ($\%SS_e$) is represented by the remaining part (white).

DISCUSSION

Using high-quality protein intensity data and a quantitative mapping approach, differences between the chromosome substitution lines were mapped to specific chromosomes or combinations thereof. Additionally, the contribution of epistasis to protein expression variation was interrogated. We have shown that protein intensities can depend on a combination of multiple allelic chromosomes. The current dataset also allows to differentiate between potentially different alleles of the protein encoding genes due to the accession specific databases available for both parental accessions. However, it is certain that the full potential of the current dataset has not yet been exhaustively explored in the current study, and there are possibilities for advancing the current analysis.

From our analysis it becomes clear that especially chromosomes 2 and 5 influence the variation in the proteome. As discussed before, this is not completely unexpected as previous research has shown how both these chromosomes also contain hotspots for eQTL and phenotypic QTLs. It is highly likely that the same genes (at least *erecta* on chromosome 2) are causal for the large variation in protein abundance as well. Similarly, on chromosome 5 there are a number of major developmental regulators present that could potentially influence the proteome as well.

The current experiment was set-up as an exploratory investigation and as a consequence it has inspired thoughts for follow up experiments and additional analyses. First, the proteomics experiment could be extended to include more CSL genotypes to make the analysis more comprehensive and complete. With additional genotypes more genetic effects can be estimated with greater accuracy. A second option would be to investigate different tissues of the same plants or the different response of the CSLs by inclusion of a treatment for the plants. This is especially interesting as the largest protein quantity differences identified in the current conditions were related to (a)biotic stress response. For several of the proteins with the largest effect QTCs experimental evidence suggests that RNA expression variation leads to different phenotypes, which would be interesting to confirm for the set of CSLs using RNAseq or qPCR techniques.

Furthermore, proteomics could be combined with other -omics technologies, like transcriptomic and metabolomics. This can determine whether the genetic effects identified for protein abundance are an effect of differential RNA expression, or whether it is dependent on protein-protein interactions, protein-turnover or post-translational modifications. Also, follow up experiments to confirm that there is a link between RNA expression, protein abundance and a specific morphological phenotype would greatly improve the confidence that the detected genetic effects have biological meaning and whether selection can occur on protein abundance.

The proteomics approach itself could also be extended by other methodologies to improve detection of more proteins. Although there is no protocol for detecting all proteins from a single extraction, the current protein extraction method is limited to detect mainly hydrophobic proteins. Generalization of the current results should therefore be limited to this subset of proteins, but information of additional proteins, by performing different extraction methods, would potentially allow a comprehensive view of protein networks and their relation with other -omics data to obtain valuable information for systems biology applications.

In contrast to obtaining a comprehensive proteome overview, a targeted proteomics approach of TQS (Triple-Quad-Spectrometry) would allow the fast detection of specific proteins of interest. Such a method allows detection and identification of proteins with shorter scan times and more precision. For instance it would be interesting to verify the epistatic effects detected for the twenty proteins, for which the interaction was identified as the most significant term in the final model. As mentioned before, for these proteins there is a clear indication for the presence of a sign epistatic effect where opposite allelic combinations lead to similar outcomes. Also according to the explained epistatic variance ($\%SS_i$) those sign epistatic interactions had a major contribution to the genetic variance ($\%SS_g$), and thus might have a biological relevance.

The current approach for epistasis detection with chromosome substitution lines combined with shotgun proteomics provides a first glimpse of the versatile use for chromosome substitution lines. We wanted to investigate a subset of the complete CSL panel as was used by Wijnen et al. to see if the use of only single and double chromosome substitutions of L*er* in the Col recurrent background would suffice for detection of different genetic effects. With this approach only a subset of all possible genetic effects can be estimated, but this is more comprehensive than is currently possible in most biparental segregating populations. Obviously, this advantage stems also from the small size of the current panel in which only a limited number of genetic effects are present, which all can be tested easily in a single model.

The ANOVA-based approach with backward elimination of non-significant terms for detecting the genetic effects in the current CSL panel suggests that a significant QTC_l can be ultimately additively introgressed (or maintained) in a Col-0 genetic background. However in a different genetic background the QTC_l might lead to a non-significant response. Mapping epistasis with a ANOVA-based approach is also allele frequency dependent. For instance the different haplotype classes CC; CL; LC and LL for the combination of chromosome 1 and 5 are present at the following frequencies: 36%-29%-21%-14%. This triggers the suggestion for an additional analysis in which only four genotypes are tested for each epistatic effect. For instance, along the lines

of the current ANOVA-based approach the main effects and interaction effect for chromosomes 4 and 5 can be tested in only the genotypes CCCCC, CCCCL, CCCLC and CCCLL, where haplotype frequency is 25% for each combination of chromosome 4 & 5. It is likely that in such an analysis not all detected interactions (and main effects) of the current study will be equally significant. However, when one is only concerned about introgressing a specific QTC_I into a parental line, this might suffice.

We identified 142 significant interaction effects for 18% of the selected proteins. Although we have attempted here to quantify the frequency at which epistasis occurs, it needs to be clear that epistasis likely is influenced at different biological levels. There is ample evidence that protein-protein interactions cause additional phenotype variation. Indeed, post-translational modifications and protein-complex forming still occur downstream of the current measurements of protein quantity. Additionally, with CSLs we provide the tools for detection of inter-chromosomal epistasis. It is highly likely there is equal if not more intra-chromosomal epistasis present.

ACKNOWLEDGEMENTS

We like to express our gratitude to G. Stunnenberg, T. Stoker and R. van Genderen of Wageningen University for technical assistance during experimental work. Additionally we are grateful to J. Cordewener and J. van den Heuvel for useful discussions. Funding: This work has been financially supported by the Netherlands Organisation for Scientific Research under grant number STW-12425 for which additional support was received from Rijk Zwaan B.V.. Additional funding was acquired via a NOW-ZonMw partnership under project number 435003019.

REFERENCES

1. I. The Arabidopsis Genome, Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* 408, 796 (2000).
2. L. Zapata *et al.*, Chromosome-level assembly of Arabidopsis thaliana Ler reveals the extent of translocation and inversion polymorphisms. *Proceedings of the National Academy of Sciences* 113, E4052-E4060 (2016).
3. J. J. B. Keurentjes *et al.*, The genetics of plant metabolism. *Nature Genetics* 38, 842 (2006).
4. N. Carreno-Quintero, H. J. Bouwmeester, J. J. B. Keurentjes, Genetic analysis of metabolome–phenotype interactions: from model to crop species. *Trends in Genetics* 29, 41-50 (2013).
5. R. W. Doerge, Mapping and analysis of quantitative trait loci in experimental populations. *Nature Reviews Genetics* 3, 43 (2002).
6. R. C. Jansen, J.-P. Nap, Genetical genomics: the added value from segregation. *Trends in Genetics* 17, 388-391 (2001).
7. F. Edfors *et al.*, Gene-specific correlation of RNA and protein levels in human cells and tissues. *Molecular Systems Biology* 12, (2016).
8. A. Ghazalpour *et al.*, Comparative Analysis of Proteome and Transcriptome Variation in Mouse. *PLOS Genetics* 7, e1001393 (2011).
9. M. Muers, Transcriptome to proteome and back to genome. *Nature Reviews Genetics* 12, 518 (2011).
10. K. Kosová, P. Vítámvás, I. T. Prášil, J. Renaut, Plant proteome changes under abiotic stress — Contribution of proteomics studies to understanding plant stress response. *Journal of Proteomics* 74, 1301-1322 (2011).
11. C. A. Schenck *et al.*, A proteomics approach identifies novel proteins involved in gravitropic signal transduction. *American Journal of Botany* 100, 194-202 (2013).
12. F. Chevalier *et al.*, Proteomic investigation of natural variation between Arabidopsis ecotypes. *PROTEOMICS* 4, 1372-1381 (2004).
13. C. L. Wijnen *et al.*, A complete chromosome substitution mapping panel reveals genome-wide epistasis in Arabidopsis. *bioRxiv*, (2018).
14. J. B. Singer *et al.*, Genetic dissection of complex traits with chromosome substitution strains of mice. *Science* 304, 445-448 (2004).
15. J. K. Belknap, Chromosome substitution strains: some quantitative considerations for genome scans and fine mapping. *Mammalian Genome* 14, 723-732 (2003).
16. H. Shao *et al.*, Genetic architecture of complex traits: Large phenotypic effects and pervasive epistasis. *Proceedings of the National Academy of Sciences* 105, 19910-19914 (2008).
17. A. Chen, Y. Liu, S. M. Williams, N. Morris, D. A. Buchner, Widespread epistasis regulates glucose homeostasis and gene expression. *PLOS Genetics* 13, e1007025 (2017).
18. W. Kruijer *et al.*, Marker-based estimation of heritability in immortal populations. *Genetics* 199, 379-398 (2015).
19. D. W. Huang, B. T. Sherman, R. A. Lempicki, Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols* 4, 44 (2008).
20. B. T. Sherman, D. W. Huang, R. A. Lempicki, Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research* 37, 1-13 (2008).
21. B. D. Rosen *et al.*, Araport: the Arabidopsis Information Portal. *Nucleic Acids Research* 43, D1003-D1009 (2014).
22. A. Vertommen, B. Panis, R. Swennen, S. C. Carpentier, Evaluation of chloroform/methanol extraction to facilitate the study of membrane proteins of non-model plants. *Planta* 231, 1113-1125 (2010).
23. A. Michalski *et al.*, Mass Spectrometry-based Proteomics Using Q Exactive, a High-performance Benchtop Quadrupole Orbitrap Mass Spectrometer. *Molecular & Cellular Proteomics* 10, (2011).
24. Y. Zhang, B. R. Fonslow, B. Shan, M.-C. Baek, J. R. Yates, Protein Analysis by Shotgun/Bottom-up Proteomics. *Chemical Reviews* 113, 2343-2394 (2013).
25. R. Magrath *et al.*, Genetics of aliphatic glucosinolates. I. Side chain elongation in Brassica napus and Arabidopsis thaliana. *Heredity* 72, 290 (1994).
26. J. Kroymann *et al.*, A Gene Controlling Variation in Arabidopsis Glucosinolate Composition Is Part of the Methionine Chain Elongation Pathway. *Plant Physiology* 127, 1077-1088 (2001).

27. S. Textor, J.-W. de Kraker, B. Hause, J. Gershenzon, J. G. Tokuhisa, MAM3 Catalyzes the Formation of All Aliphatic Glucosinolate Chain Lengths in Arabidopsis. *Plant Physiology* 144, 60-71 (2007).
28. J. Kroymann, S. Donnerhacke, D. Schnabelrauch, T. Mitchell-Olds, Evolutionary dynamics of an Arabidopsis insect resistance quantitative trait locus. *Proceedings of the National Academy of Sciences* 100, 14587-14592 (2003).
29. D. Winter *et al.*, An "Electronic Fluorescent Pictograph" Browser for Exploring and Analyzing Large-Scale Biological Data Sets. *PLOS ONE* 2, e718 (2007).
30. B. Field *et al.*, Glucosinolate and Amino Acid Biosynthesis in Arabidopsis. *Plant Physiology* 135, 828-839 (2004).
31. B. G. Hansen *et al.*, A Novel 2-Oxoacid-Dependent Dioxygenase Involved in the Formation of the Goiterogenic 2-Hydroxybut-3-enyl Glucosinolate and Generalist Insect Resistance in Arabidopsis. *Plant Physiology* 148, 2096-2108 (2008).
32. G. K. Pattanayak *et al.*, ACCELERATED CELL DEATH 2 suppresses mitochondrial oxidative bursts and modulates cell death in Arabidopsis. *Plant J* 69, 589-600 (2012).
33. N. Yao, J. T. Greenberg, Arabidopsis ACCELERATED CELL DEATH2 Modulates Programmed Cell Death. *The Plant Cell* 18, 397-411 (2006).
34. H. Cao, J. Glazebrook, J. D. Clarke, S. Volko, X. Dong, The Arabidopsis NPR1 Gene That Controls Systemic Acquired Resistance Encodes a Novel Protein Containing Ankyrin Repeats. *Cell* 88, 57-63 (1997).
35. S. H. Spoel *et al.*, NPR1 modulates cross-talk between salicylate- and jasmonate-dependent defense pathways through a novel function in the cytosol. *Plant Cell* 15, 760-770 (2003).
36. E. Bell, R. A. Creelman, J. E. Mullet, A chloroplast lipoxygenase is required for wound-induced jasmonic acid accumulation in Arabidopsis. *Proceedings of the National Academy of Sciences* 92, 8675-8679 (1995).
37. J. Fu *et al.*, System-wide molecular evidence for phenotypic buffering in Arabidopsis. *Nature Genetics* 41, 166 (2009).
38. K. U. Torii *et al.*, The Arabidopsis ERECTA gene encodes a putative receptor protein kinase with extracellular leucine-rich repeats. *The Plant Cell* 8, 735-746 (1996).
39. I. R. Terpstra, L. B. Snoek, J. J. B. Keurentjes, A. J. M. Peeters, G. Van den Ackerveken, Regulatory network identification by genetical genomics: Signaling downstream of the arabidopsis receptor-like kinase ERECTA. *Plant Physiology* 154, 1067-1078 (2010).
40. M. van Zanten, L. B. Snoek, M. C. G. Proveniers, A. J. M. Peeters, The many functions of ERECTA. *Trends in Plant Science* 14, 214-218 (2009).
41. G. Jander, J. Cui, B. Nhan, N. E. Pierce, F. M. Ausubel, The TASTY Locus on Chromosome 1 of Arabidopsis Affects Feeding of the Insect Herbivore Trichoplusia ni. *Plant Physiology* 126, 890-898 (2001).
42. M. d. T. Zabala *et al.*, Characterisation of recombinant epithiospecifier protein and its over-expression in Arabidopsis thaliana. *Phytochemistry* 66, 859-867 (2005).
43. X.-L. Liu, H.-D. Yu, Y. Guan, J.-K. Li, F.-Q. Guo, Carbonylation and Loss-of-Function Analyses of SBPase Reveal Its Metabolic Interface Role in Oxidative Stress, Carbon Assimilation, and Multiple Aspects of Growth and Development in Arabidopsis. *Molecular Plant* 5, 1082-1099 (2012).
44. C. Alonso-Blanco *et al.*, 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell* 166, 481-491 (2016).
45. Z. Zhang, J. A. Ober, D. J. Kliebenstein, The gene controlling the quantitative trait locus EPITHIOSPECIFIER MODIFIER1 alters glucosinolate hydrolysis and insect resistance in arabidopsis. *The Plant Cell* 18, 1524-1536 (2006).
46. M. Burow *et al.*, ESP and ESM1 mediate indol-3-acetonitrile production from indol-3-ylmethyl glucosinolate in Arabidopsis. *Phytochemistry* 69, 663-671 (2008).

SUPPLEMENTARY MATERIALS

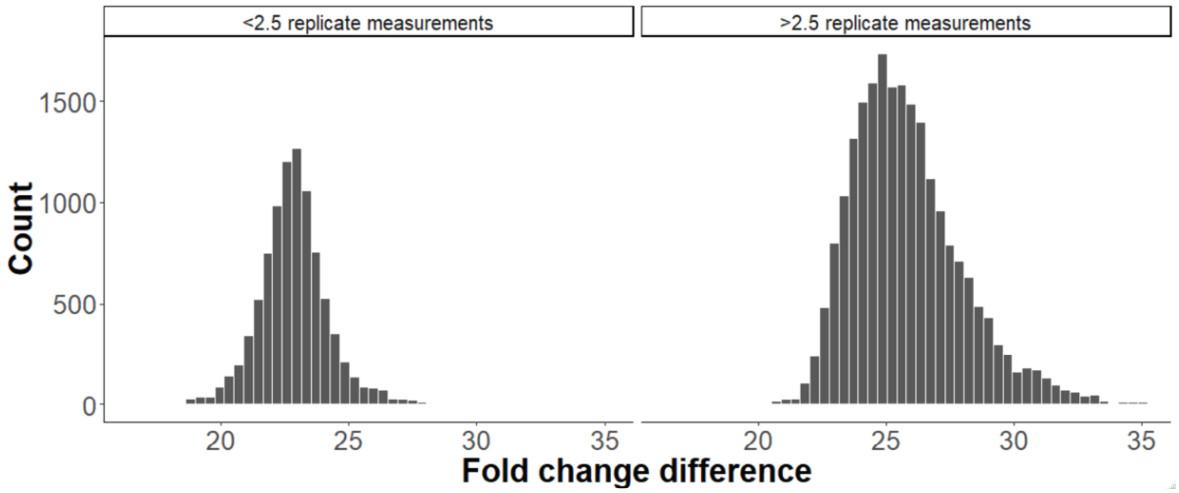


FIGURE S1 | The intensity influences the number of detections per genotype. The difference in distribution of the average log2-protein intensity values per genotype of each protein. The left histogram shows proteins with less than 2.5 detections per genotype on average (22.8 ± s.d. 1.37), the right shows proteins with more than 2.5 detections per genotype on average (25.8 ± s.d. 2.16).

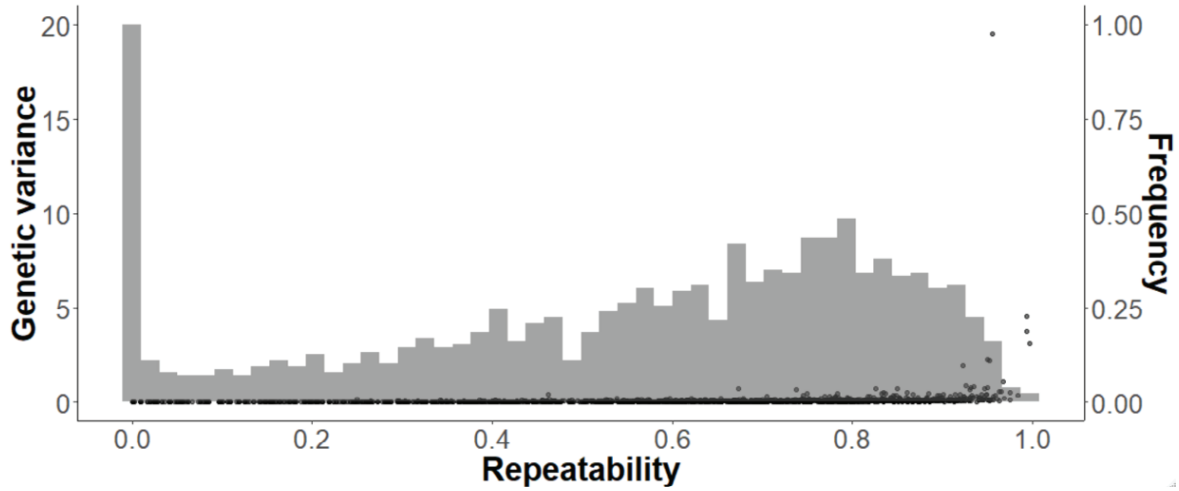


FIGURE S2 | Heritability and genetic variance of the 1,468 proteins. The dots represent proteins according to their respective repeatability (x-axis) and genetic variance (left y-axis). The histogram shows the distribution of the heritabilities for the quantitative proteins (right y-axis). While there were many proteins with a large (>0.5) heritability, the genetic variance was in general low.

Chapter 5

Transient crossover reduction by virus-induced gene silencing enables efficient reverse breeding

Vanesa Calvo-Baltanás[1], Cris L. Wijnen[1], Nina Lukhovitskaya[2#a], C. Bastiaan de Snoo[3], Linus Hohenwarter[4], Joost J. B. Keurentjes[1], Hans de Jong[1], Arp Schnittger[4] and Erik Wijnker[1].

[1] Laboratory of Genetics, Wageningen University & Research, Wageningen, the Netherlands.
[2] Institut de Biologie Moléculaire des Plantes, Centre National de la Recherche Scientifique, Université de Strasbourg, Strasbourg, France.
[3] Rijk Zwaan R&D Fijnaart, Fijnaart, the Netherlands.
[4] Department of Developmental Biology, Biocenter Klein Flottbek, University of Hamburg, Hamburg, Germany.
[#a] Current address: Division of Virology, Department of Pathology, University of Cambridge, Tennis Court Rd, Cambridge, CB2 1QP, UK.

This chapter has been submitted for publication and is online available via BioRxiv: biorxiv.org/content/10.1101/459016v1

ABSTRACT

Reverse breeding is the opposite of plant hybridization: a method to extract parental lines from a hybrid. Reverse breeding enables the development of new F1 hybrid varieties without having prior access to homozygous breeding lines. It also can be used to efficiently generate chromosome substitution lines (CSL). For successful reverse breeding, the heterozygotes' chromosomes must be divided over two haploid complements. This is achieved by suppression of meiotic crossover recombination and the subsequent production of doubled haploids. We here show two innovations that allow efficient reverse breeding. Firstly, we demonstrate that transgene-free offspring with a desired genetic-make up can be produced by using virus-induced gene silencing to downregulate meiotic crossovers. Secondly, our experiments suggest that incomplete rather CO suppression enables reverse breeding to be efficiently applied in other species.

INTRODUCTION

Heterozygous F1 hybrids are among the highest producing crop varieties (1) and result from intercrossing homozygous parental lines. Existing hybrids are usually further improved through the introgression of new alleles into their parental lines. In an alternative approach, large numbers of new and potentially better heterozygous genotypes could be generated in outcrossing populations, for example by intercrossing different commercially available heterotic hybrids and selecting the best performing heterozygotes in their offspring. However, this potential is rarely, if ever exploited, because unique heterozygotes selected from outcrossing populations cannot be maintained: when they set seed, their unique allele combinations are lost through meiotic recombination. This restriction can be overcome by reverse breeding, in which new parental lines for any heterozygote can be *post-hoc* generated from the selected heterozygote itself (2–4) (Fig 1). By obtaining its parental lines, a heterozygote can be recreated as F1 hybrid.

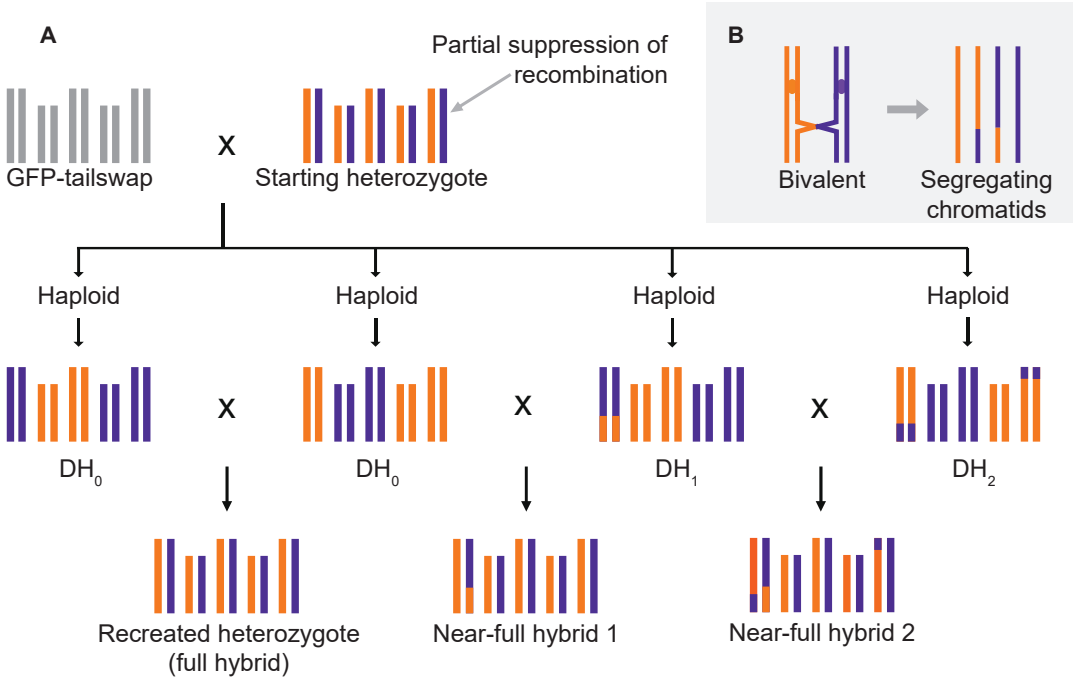


FIGURE 1 | Reverse breeding (heterozygote reconstruction) through partial crossover (CO) suppression in Arabidopsis. Panel A: a starting Arabidopsis heterozygote (top) is selected for which parental lines are to be made. Five chromosome pairs are shown, with homologs in orange and purple. Meiotic COs are partially suppressed in this heterozygote, after which gametes are formed that are grown into haploid, and subsequently doubled haploid (DH) offspring as shown in the middle row having 0, 1 or 2 COs (DH_0, DH_1 and DH_2 respectively). Intercrossing complementing DH_0 (left) recreates the heterozygote as a full hybrid (bottom row, left), an approach similar as described by Wijnker et al., 2012 (4). Intercrossing DH_0 with DH_1 (middle) or DH_1 with DH_2 (right) generates near-full hybrids 1 and 2, which have small homozygous genomic regions. Note that in the cross of DH_1 with DH_2 chromosome 1 is largely heterozygous, since the parental lines complement one another in the distal chromosome region. Panel B: Recombinant but also non-recombinant chromatids segregate in the presence of a CO. Detail of a bivalent pair with one meiotic CO is shown (left). Only two of the four resulting chromatids are recombinant (right).

A proof of concept study (4) showed the feasibility of reverse breeding in an *Arabidopsis thaliana* hybrid. This was achieved by the complete knock-down of meiotic CO formation in a F1 hybrid using a dominantly acting RNAi transgene targeting the essential meiotic recombinase DISRUPTED MEIOTIC CDNA 1 (DMC1). Without COs, non-recombinant chromosomes segregate to gametes. These gametes were regenerated as haploid plants, and self-fertilized to give rise to homozygous diploid lines (doubled haploids; DH) from which complementing parental lines were selected and crossed to reconstitute the starting heterozygote (Fig 1A). In short, reverse breeding requires the consecutive suppression of recombination and the conversion of resulting gametes to DH offspring.

Alternatively, reverse breeding can serve as an efficient technique to generate chromosome substitution lines (CSLs). In this case, two parental lines are crossed to give rise to an F1. In the absence of crossover recombination during meiosis and due to the random segregation of chromosomes, different combinations of homozygous parental chromosomes are transmitted to DH offspring, leading to the generation of chromosome substitution lines (CSLs) (3–5). CSLs have been shown to be outstanding tools for mapping QTL in mice, and allow for the systematic detection of epistasis (non-allelic interactions) (6–8) The generation of a complete CSL population in Arabidopsis allowed for the systematic detection of two-way or three-way interactions for different traits (5).

Since the translation of this technique to crops may be challenging, we here set out to overcome two major drawbacks of the original approach. Firstly, the use of a transgene to suppress CO formation in a heterozygote is impractical. Stable transformation of a selected heterozygote can be complex and a transgene that dominantly compromises fertility renders half of the offspring (the genotypes carrying the construct) useless for further breeding. We asked if virus-induced gene silencing (VIGS) could be used to transiently suppress meiotic CO formation in a hybrid (9–11) and whether gametes resulting from VIGS-modified meiosis can be used to generate offspring of desired genotypic composition.

Secondly, CO formation is indispensable for chromosome segregation in plants. Without COs, homologs segregate randomly (as univalents) at anaphase I. This causes aneuploidy in gametes and semi-sterility. Viable haploid gametes can still be formed in the absence of COs, when the homologues of each chromosome pair by random chance segregate to opposite poles (4). The probability of regular disjunction is a function of the chromosome number of the plant (3) (n): $P_{(\text{balanced segregation})} = 1/2^n$. The more chromosome pairs, the lower the probability of viable gamete formation, and the lower the chance of obtaining parental lines. In Arabidopsis about $1/2^5$ = 3% of meiotic events generates viable spores in the absence of COs.

The suppression of CO formation enriches for the segregation of non-recombinant chromosomes to gametes, but complete CO suppression is not essential. Gametes carrying exclusively non-recombinant chromosomes will occasionally be formed in wild-type meiosis (Fig 1B) although they are usually rare, especially when the chromosome number is high (S1-S6 Tables). A reduction of CO, rather than complete CO suppression, might present a favourable intermediate approach to enrich for viable gametes carrying only or mostly, non-recombinant chromosomes (3). The presence of parallel pathways that lead to CO formation in plants allows theoretically for fine-tuning CO rates (12). Mutants of *MUTS HOMOLOGUE5 (MSH5)* show about 87% reduction in COs in Arabidopsis (13) and we therefore targeted *MSH5* using a VIGS construct to reduce CO formation.

RESULTS

Downregulation of MSH5 using VIGS changes the genetic composition of offspring

The efficiency of VIGS to downregulate *MSH5* was assessed by inoculating plants at the five-leaf stage with a VIGS vector (TRV2-*AtMSH5*) and evaluating meiotic progression. *MSH5* knocked-down plants exhibited high levels of aborted pollen about three weeks after inoculation and siliques that failed to elongate (Fig 2) consistent with a *msh5* mutant phenotype (14). Chromosome spreads of late meiotic cell complements confirmed the mis-segregation of chromosomes during meiosis (Fig 2). High levels of pollen abortion were typically observed for three to four consecutive days on open flowers, which amounts to about six to eight flowers per plant. Thereafter the plants reverted to showing a wild-type phenotype, in which flowers exhibit viable pollen and form long siliques.

To evaluate the feasibility of breeding with gametes resulting from VIGS-mediated reduction of recombination, we inoculated an F1 (Landsberg *erecta* x Columbia) with TRV2-*AtMSH5*. Once the flowers showed a high fraction of aborted pollen, they were crossed to *GFP-tailswap*, a haploid inducer line for Arabidopsis (15). Haploid offspring were obtained and self-fertilized to give rise to 111 DH offspring that were genotyped for 42 markers evenly spaced over the genome (S1 Fig).

Among the 111 offspring we identified 24 DHs (20 different genotypes) that carry only non-recombinant chromosomes (S1 Dataset). These lines, which are essentially chromosome substitution lines (CSLs), are henceforth referred to as DH_0 to differentiate these offspring from other DH in the following section. The population is significantly enriched for DH_0 lines in comparison to a previously published wild-type DH population (Fig 3) (Kolmogorov-Smirnov test; α=0.01). Among these 20

DH_0 genotypes we identified six complementing parental pairs that, when crossed, recreate the starting hybrid (S1 Dataset). All DH offspring in our population developed normally and were fully fertile. This shows that VIGS can transiently modify meiotic recombination in wild-type hybrids and change the genetic make-up of offspring derived from a hybrid.

Recreation of hybrid genotypes and phenotypes using reverse breeding offspring

For the exact recreation of a heterozygous *genotype*, complementing DH_0 are required, but in practice the recreation of the hybrid *phenotype* will be the ultimate goal. The use of a DH_1 (i.e. a DH with one recombinant chromosome) in a cross to recreate a hybrid, leads to a decrease of heterozygosity (hereafter DOH) in the reconstituted hybrid distal to the CO position (Fig 1). We hypothesized that only in case DOH negatively affects the hybrid phenotype, it is of concern for reverse breeding. In our offspring we identified 19 DH_1 and 12 DH_2 with one and two COs per genome respectively, with the remainder of 58 DH having three to eight COs, which is in the range of wild-type meiosis and likely result from incomplete penetrance of VIGS (S1 Dataset). Depending on CO positions in the DH_1 and DH_2 offspring, we noted the possibility of identifying four additional parental pairs in which a near-full hybrid would show a DOH less than 2.5% of the total genome length. Seven parental

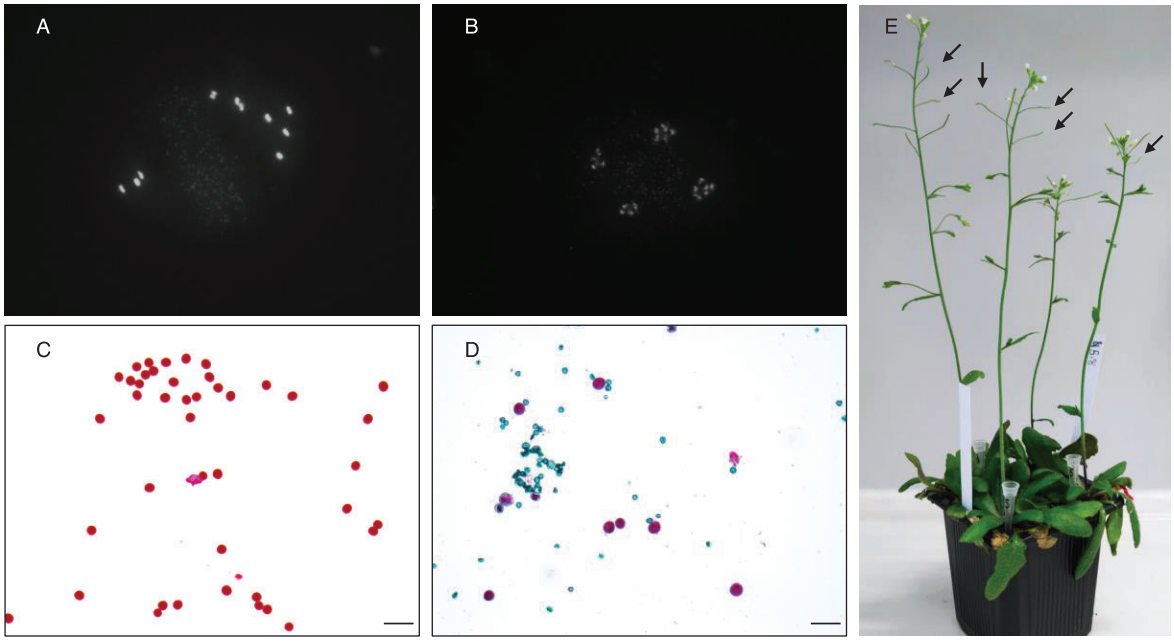


FIGURE 2 | Inoculation with TRV2-*AtMSH5* induces a phenocopy of a *msh5* mutant phenotype in *Arabidopsis thaliana* hybrids during meiosis, in pollen phenotype and plant fertility. A and B depict the result of random chromosome segregation at metaphase II (A) and tetrad stage (B) (same magnification). Pollen of a non-inoculated control plant and an inoculated plant with TRV2-*AtMSH5* is shown in (C) and (D), scale bar 50 µm. The result of a *MSH5* knock-down leads to high pollen abortion in (D). *MSH5* knocked-down plants display short siliques indicated with black arrows (E).

pairs would give rise to near-full hybrids in which DOH is less than 5%. Only near-full hybrids with one CO show less than 2.5 % DOH. In one parental pair (DH_1 line 44 x DH_2 line 41) COs on the same chromosome arm partly compensate, similar to the DH_1 x DH_2 cross illustrated in Fig 1, generating a near-full hybrid with a DOH of 4.2%.

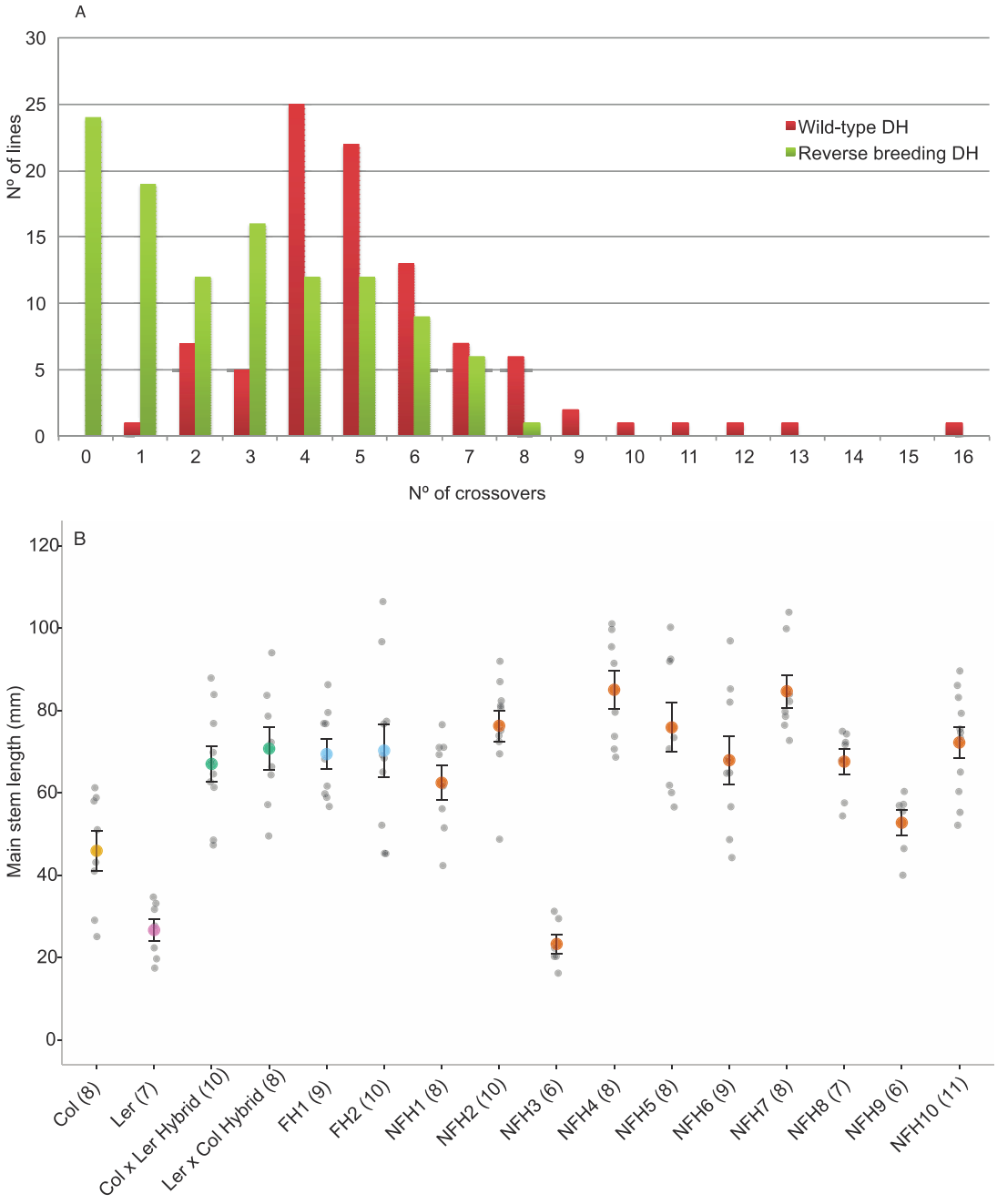


FIGURE 3 | Crossover (CO) distributions in reverse breeding- and wild-type DH offspring (top panel) and comparison of hybrid phenotypes (lower panel). Top panel (A) shows the observed CO number in wild-type DH offspring (in red) and reverse breeding DH offspring (in green). Note that reverse breeding DH offspring are enriched for DH having 0 and 1 COs. Lower panel (B) shows box-and-whisker plots for main stem length at the moment of flowering of parental lines, full hybrids (FH) and near-full hybrids (NFH). Parental lines Col-0 and L*er* (shown in yellow and pink respectively), Col-0 x L*er* reciprocal hybrids (green), full hybrids (blue) and near-full hybrids (orange). Genotypes of near full hybrids are presented in the supplementary file.

The phenotypic impact of DOH can be explored experimentally. We therefore intercrossed DH lines to create near-full hybrids with increasing levels of DOH ranging from 1.28% - 32.1% (Fig 3B and S2). These were grown together with the starting heterozygote and full hybrids (recreated heterozygotes) and compared standard growth parameters: flowering time, main stem length, rosette diameter and dry weight at flowering time. No significant differences were found between the starting hybrid and the full hybrids (one-way ANOVA; FT p-value = 0.3015; MSL p-value = 0.9347; RD p-value = 0.8655; DW p-value = 0.2697; Fig 2B and S2). Also, no significant differences between the full hybrid and the near-full hybrids were found, with the exception of one: a near-full hybrid that has a similar short stem length as one of its parental lines (Fig 3B), which is likely caused by homozygosity of the main effect *erecta* locus that is homozygous in this specific hybrid (16). Interestingly, NFH4 (DOH=32,07%), NFH5 (DOH=28,08%), NFH7 (DOH=31,03%), NFH9 (DOH=14,25%) and NFH10 (DOH=21,46%) that show the highest DOH (S1 Dataset) do not display negative heterosis for the trait main stem length; instead slightly higher mean values are observed compared to the full-hybrids (Fig 3).

DISCUSSION

Our experiments showed the feasibility of using VIGS to transiently modify meiotic recombination to change the genetic composition of gametes. The possibility of generating offspring from these modified gametes opens up routes to also alter other meiotic processes to change the genetic composition of offspring. It was previously reported that VIGS could be used to silence meiotic genes and induce a meiotic mutant phenotype in wheat (17,18). This suggests that VIGS-mediated silencing of meiotic genes can be used to develop breeding strategies in other species too.

In comparison to a previously published proof-of-concept for reverse breeding (4,19), the execution of reverse breeding is here greatly simplified and improved. Firstly, the previous experiments did in fact not recreate a wildtype hybrid, but recreated a specially designed transgenic achiasmatic hybrid that expressed a dominant acting RNAi transgene to suppress crossover formation. Here the application of VIGS allowed the direct suppression of CO formation in a wildtype heterozygote. Where the experiment by Wijnker et al., (2012) required six generations (three to create a transgenic hybrid, and three to recreate it), the current experiment required just three generations to recreate a heterozygote as F1 seeds (Fig 1). Furthermore, all recovered offspring in this experiment are transgene-free and fertile, while in the previous set-up half of the offspring were transgenic and semi-sterile, which implies two-fold increase of efficiency.

Our experiments also show that CO formation during meiosis can be adjusted to favourable levels, by targeting *MSH5* rather than *DMC1* as was previously done. The lower the CO number, the more DH_0 and DH_1 occur in the offspring (S1-S6 Tables), but also the higher the level of gamete abortion. Depending on ones' interest in obtaining DH_0 and DH_1, the optimal CO rate can be calculated to balance one against the other. Especially in species with higher chromosome numbers, such considerations matter. In a species with ten chromosome pairs, in which a typical bivalent has two COs, complete CO suppression generates 100% DH_0 offspring but results in just 0.10% of spore viability. Reducing COs by 75% -from 20 to five COs per meiosis- would increase spore viability 32 fold to 3%, since then only five rather than ten univalent pairs segregate. Of those offspring, 5.6% are DH_0 (S3 Table). This is low in comparison to complete CO suppression, but it is a substantial 60.000 fold increase in comparison to wild-type meiosis. Likewise, chances for obtaining DH_1 in its offspring are 18.8%, equal to about 10.000 fold increase. S1- S6 Tables give expected DH_0 and DH_1 numbers at different levels of CO suppression and different chromosome numbers. Such calculations will help to determine the best possible approach for other species than Arabidopsis.

The results obtained further illustrate that DOH do not necessarily negatively impacts hybrid performance. It is possible to estimate the expected DOH in near-full hybrids resulting from a single CO. Arabidopsis has five linkage groups (chromosomes). One CO recombines one linkage group ($1/5^{th}$) and this CO exchanges anything between zero and half of the linkage group, which averages at $1/4^{th}$ of the linkage group (typically half a chromosome arm). Expected DOH caused by a single CO thus equals on average (1/4*1/5=) 5% of the total linkage map length. Of the ten near-full hybrids (with one CO) that we can produce, five have a DOH less than 5% in Mbp, exactly as predicted as the Arabidopsis genetic map correlates well with the physical chromosome length. The more chromosomes a species has, the lower the relative DOH resulting from one CO. In a species with ten chromosome pairs (e.g. maize) one CO causes a DOH of 2.5%. This decreases even further when, as in many species, COs locate relatively distal on chromosomes. Under such a scenario, not only DH_0, but also DH_1 and DH_2 may prove worthy parental lines, provided that resulting near-full hybrids are phenotyped to assess their performance.

Reverse breeding is a method to generate parental lines for heterozygotes, but it also provides a way to generate populations of DH_0, also known as chromosome substitution lines (4). Due to the low number of segregating loci (chromosomes), such populations are near unparalleled tools to identify QTL and map epistatic interactions (5–8). Since in mixed DH_0/DH_1 populations the number of segregating loci remains near minimal, the detection power of QTLs and epistasis is unlikely to

decrease much. Reverse breeding strategies are therefore a way towards detection and mapping complex interactions. The recent identification of class I COs meiotic mutants in crops like maize, *Brassica,* rice or barley present further interesting opportunities for developing reverse breeding strategies based on partial crossover suppression (20–23).

The application of reverse breeding can provide a route to the quick generation of CSLs (Wijnker et al., 2012), and it has been shown that all 32 possible chromosome combinations can be obtained for Arabidopsis with the use of a dominant transgene. For the detection of QTL, single CLSs (sCLSs) in which one chromosome is introgressed in a recurrent background the minimal population comprises as many lines as the basic chromosome number of the species of interest. By also generating double CSLs (with two introgressed chromosomes), simple (two-way) epistatic interactions can also be detected. Transient suppression of CO recombination using VIGS facilitates the generation of CSLs because no construct segregates in the offspring, and because COs can be suppressed in any susceptible hybrid genotype.

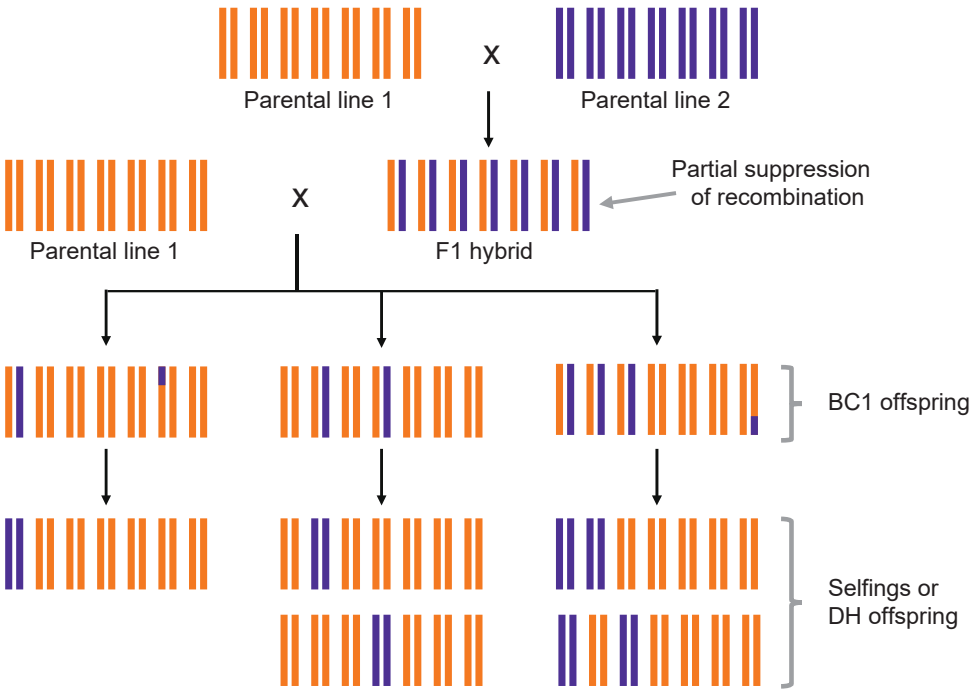


FIGURE 4 | Generation of single and double chromosome substitution lines by reduced recombination in a species with 7 chromosome pairs. Parental lines are crossed to give rise to an F1 hybrid. Crossover recombination is reduced (or may be completely suppressed) in the F1 hybrid, after which it is backcrossed to a parental line (second row). An example here is given of a backcross to parental line 1. The BC1 offspring will contain mainly non-recombinant chromosomes (third row), of which the number of non-recombinant, heterozygous chromosomes in the BC1 follows a binomial distribution. The partially heterozygous BC1 offspring can be left to self-fertilize, or be used for the production of doubled haploids (DH) to give rise to homozygous chromosome substitution lines (CSLs). The likeliness of obtaining homozygous CSLs among DH offspring or offspring obtained through self-fertilization can be increased when also the BC1 are subjected to (partial-) CO suppression.

In our experiment we obtained 20 different CSLs, but by chance only four of these were sCSLs. To complete the set of sCSLs for our starting parental lines, we could backcross DH line 4 to Col-0, and Lines 12 and 23 to L*er* (S1 Dataset) to give rise to hybrids that are heterozygous for two chromosomes, from which the missing single CSLs could be obtained with relative ease by the production of DH from these partial hybrids. Using VIGS, it is also possible to design relatively simple backcross designs, that would enable to obtain these lines in more targeted ways. In Fig 4 such a simple backcross design is shown, in which an F1 in which recombination is (partially) suppressed is backcrossed to one of its parents. This gives rise to BC1 offspring in which only some chromosome pairs are heterozygous. The number of heterozygous chromosome pairs follows a classical binomial distribution. From such a backcross population one could select partial hybrids with 1, 2 or 3 heterozygous chromosome pairs, from which single and double CSLs can be obtained through self-fertilization or by the production of DHs. The chances of obtaining the required single or double CSLs depends of the (expected) CO frequency in the partial hybrid. In such a breeding scheme, the researcher will be able to calculate as to whether it is convenient to also suppress crossover recombination in the BC1 as well in order to increase the chance of obtaining the desired offspring.

MATERIALS AND METHODS

Plant material and growth

Arabidopsis thaliana plants used in crosses and for VIGS inoculation were grown in potting soil in growth chambers (Percival), 21º/18º C, 16H / 8H light cycle and 60%-50% relative humidity. Haploid offspring were grown under similar conditions in a greenhouse. For phenotyping, seeds of DH offspring, reconstituted full hybrids and near-full hybrids were vernalized by sowing on wet filter paper and placed them for several days in the dark at 4°C for four days to ensure uniform germination. Plants were grown on 4x4 cm Rockwool blocks and watered with a flooding system with a Hyponex nutrient solution three times per week in a randomized block design with five blocks and two replicates per genotype in each block. Climate chamber conditions were set to 16h/8h and 20/18˚C day/night cycle, light was set to 125 µmolm-2s-1 and there was 70% relative humidity.

Plasmid construction and *Agrobacterium* inoculation

Two *MSH5* cDNA regions were amplified using primers to which *BamH*I (forward) and *Xba*I (reverse) restriction sites were added. The MSH5_F1/R1 and MSH5_F2/R2 primer pairs give fragments of 242 bp and 254 bp respectively, and were used to generate the TRV2-*AtMSH5* and TRV2-*AtMSH5*_2 constructs. Both PCR products were introduced individually into the vector TRV2 (pYL156) (24) following a classical digestion-ligation reaction. After sequence verification, the TRV2-*AtMSH5* and TRV2-*AtMSH5*_2 vectors were transformed into *Agrobacterium tumefaciens GV3101* (*pMP90*) strain. The incubation and inoculation protocol was executed as described in Nimchuk *et al*., 2000 (25). Plant inoculation was done by leaf-infiltration (26) of TRV2-*AtMSH5* in combination with TRV1 (pYL192) (24) or TRV2-*AtPDS* in combination with pTRV1 in a 1:1 ratio. TRV2-*AtMSH5* and TRV2-*AtMSH5*_2 induced similar pollen phenotypes in inoculated plants, after which only TRV2-*AtMSH5* was used for further experiments.

Primers used:
MSH5_F1 5′- CAGGATCCAAGCCATCGATCATTTACGC -3′
MSH5_R1 5′- CATCTAGAACTTGGACTTCACTGCCCAC -3′
MSH5_F2 5´- CAGGATCCAAGCCATCGATCATTTACGC-3´
MSH5_R2 5´- CATCTAGAACTTGGACTTCACTGCCCAC -3´

Selection of TRV-*AtMSH5* inoculated plants for pollen phenotyping

A total of 109 plants were inoculated with TRV2-*AtMSH5* in three consecutive experiments (52+42+15). Three non-inoculated plants were grown as wild-type controls in every batch as well as three to four plants in each batch that were inoculated with TRV2-*AtPDS* to silence *PDS* as positive control (10). To evaluate a successful knock-down of *MSH5*, we assessed pollen viability in flowers that opened three weeks post-inoculation and the two consecutive weeks. One anther was removed from each flower and placed on a slide with a drop of a modified Alexander stain (27) to observe pollen viability. Pollen produced by wild-type plants remained viable throughout the test periods. The number of affected flowers was not consistent. Within an inflorescence, flowers with high levels of pollen abortion usually appear consecutively, and a semi-sterile phenotype was present for up to six consecutive days after the first sterile flowers appeared.

DH production

To produce doubled haploids, F1 hybrid plants of L*er* x Col-0 plants were inoculated with TRV2-*AtMSH5* as described above. Once flowers appeared, pollen of flowers displaying high levels of dead pollen were crossed to the inducer line *GFP-tailswap* (15). Of the three consecutively grown batches 27, 19 and 15 plants were used. Other plants did not show a semi-sterile phenotype. From these plants we used 132, 77 and 60 flowers for pollination of *GFP-tailswap*. Haploid selection was done as described in Wijnker *et al*., 2014(19). Among the 369 offspring we identified 113 haploid offspring. For 111 of these we obtained DH seeds.

Phenotypical analysis of (near-)full hybrids

At the moment of flowering, flowering time (FT) was recorded and main stem length (MSL), rosette diameter (RD) and dry weight (DW) were measured for each plant. Phenotypic data was corrected for spatial trends and block effects with the SpATS R package, and the resulting spatial corrected raw data was used for further analyses. To establish whether the intercrosses of the DH0 resulted in different full hybrids, these were compared with the parental wild-type F1 using one-way ANOVA.

To assess the performance of the NFH in comparison with the FH, a Dunnett test was conducted in which line FH2 was used as a control line.

Cytology

F1 hybrid flower buds were sampled 18 days post-inoculation. The inflorescences were incubated in Carnoy: a 3:1 mix of glacial acetic acid (HAc) and 99,8% EtOH and kept overnight at 4 ºC. Inflorescences were then washed twice with 70% EtOH (in water) and stored at 4º C. Meiotic chromosome spreads were made as previously described in Ross *et al*. (1996) (28), stained with DAPI and analyzed using a Zeiss microscope equipped with epifluorescence optics.

Calculations of expected frequencies of DH_0 and DH_1.

To calculate the expected number of DH_0 and DH_1 in S1-S6 Tables, the expected number of non-recombinant and recombinant chromatids was first determined for one chromosome (i.e. the case in which the haploid chromosome number equals 1). If α is the number of COs per bivalent, then the chance of recovering a non-recombinant chromatid in a spore, and hence the chance of recovering a DH_0, equals $P_{(DH0)}=(1/2)^{\alpha}$. The chance of finding a chromatid with one CO (and hence recovering a

DH_1) equals $P_{(DH1)}=\alpha(1/2)^{\alpha}$. For higher haploid chromosome numbers (n), the expected number of non-recombinant chromatids equal $P_{(DH0)}=(1/2)^{\alpha n}$ and $P_{(DH1)}= \alpha n(1/2)^{\alpha n}$ For CO numbers 1, $P_{(DH0)}=1-\alpha/2$ and $P_{(DH1)}=\alpha/2$. For higher haploid chromosome numbers $P_{(DH0)}=(1-\alpha/2)^{n}$ and $P_{(DH1)}=n\alpha/2(1-\alpha/2)^{(n-1)}$.

ACKNOWLEDGEMENTS

This research was supported by the Netherlands the Organization for Scientific Research (NWO) through number STW-14389 (E.W.) and the European Community (EC) though the Marie-Curie Initial Training Network "COMREC", project 606956 funded under FP7-PEOPLE (V.C.-B.). We thank Cilia Lelivelt (Rijk Zwaan, Fijnaart, Netherlands) for her support in processing genotyping samples and Bas Zwaan (Wageningen University, Netherlands) for moments of reflection during our research. We thank Laurens Deurhof (Wageningen University, Netherlands) and Shinichiro Komaki (NAIST, Japan) for their support and help during the experiments.

REFERENCES

1. Schnable PS, Springer NM. Progress toward understanding heterosis in crop plants. Annu Rev Plant Biol. 2013;64(1):71–88.
2. Link W, Albrecht E M. An approach to the genetic improvement of clonal cultivars via backcrossing. Crop Sci. 1995;35(931).
3. Dirks R, Van Dun K, De Snoo CB, Van Den Berg M, Lelivelt CLC, Voermans W, et al. Reverse breeding: A novel breeding approach based on engineered meiosis. Plant Biotechnology Journal. 2009. p. 837–45.
4. Wijnker E, Van Dun K, De Snoo CB, Lelivelt CLC, Keurentjes JJB, Naharudin NS, et al. Reverse breeding in *Arabidopsis thaliana* generates homozygous parental lines from a heterozygous plant. Nat Genet. 2012;44(4):467–70.
5. Wijnen CL, Botet R, van de Belt J, Deurhof L, de Jong H, de Snoo BCB, et al. A complete chromosome substitution mapping panel reveals genome-wide epistasis in Arabidopsis. bioRxiv. 2018 Jan 1;
6. Nadeau JH, Singer JB, Matin A, Lander ES. Analysing complex genetic traits with chromosome substitution strains. Nat Genet. 2000;24(3):221–5.
7. Singer JB, Hill AE, Burrage LC, Olszens KR, Song J, Justice M, et al. Genetic dissection of complex traits with chromosome substitution strains of mice. Science. 2004;304(16).
8. Spiezio SH, Takada T, Shiroishi T, Nadeau JH. Genetic divergence and the genetic architecture of complex traits in chromosome substitution strains of mice. BMC Genet. 2012;13(38).
9. Burch-Smith TM, Anderson JC, Martin GB, Dinesh-Kumar SP. Applications and advantages of virus-induced gene silencing for gene function studies in plants. Plant J. 2004;39(5):734–46.
10. Burch-Smith TM, Schiff M, Liu Y, Dinesh-Kumar SP. Efficient virus-induced gene silencing in Arabidopsis[1]. Plant Physiol. 2006;142(1):21–7.
11. Ratcliff F, Martin-Hernandez AM, Baulcombe DC. Tobacco rattle virus as a vector for analysis of gene function by silencing. Plant J. 2001;25(2):237–45.
12. Lambing C, Franklin FCH, Wang C-JR. Understanding and manipulating meiotic recombination in plants. Plant Physiol. 2017;173(3):1530–42.
13. Higgins JD, Vignard J, Mercier R, Pugh AG, Franklin FCH, Jones GH. AtMSH5 partners AtMSH4 in the class I meiotic crossover pathway in *Arabidopsis thaliana,* but is not required for synapsis. Plant J. 2008;55(1):28–39.
14. Lu X, Liu X, An L, Zhang W, Sun J, Pei H, et al. The Arabidopsis MutS homolog AtMSH5 is required for normal meiosis. Cell Res. 2008;18(5):589–99.
15. Ravi M, Chan SWL. Haploid plants produced by centromere-mediated genome elimination. Nature. 2010;464(7288):615–8.
16. Stinchcombe JR, Weinig C, Heath KD, Brock MT, Schmitt J. Polymorphic genes of major effect: Consequences for variation, selection and evolution in *Arabidopsis thaliana*. Genetics. 2009;182(3):911–22.
17. Bennypaul HS, Mutti JS, Rustgi S, Kumar N, Okubara PA, Gill KS. Virus-induced gene silencing (VIGS) of genes expressed in root, leaf, and meiotic tissues of wheat. Funct Integr Genomics. 2012;12(1):143–56.
18. Bhullar R, Nagarajan R, Bennypaul H, Sidhu GK, Sidhu G, Rustgi S, et al. Silencing of a metaphase I-specific gene results in a phenotype similar to that of the *Pairing homeologous 1* (*Ph1*) gene mutations. Proc Natl Acad Sci. 2014;111(39):14187–92.
19. Wijnker E, Deurhof L, Van De Belt J, De Snoo CB, Blankestijn H, Becker F, et al. Hybrid recreation by reverse breeding in *Arabidopsis thaliana*. Nat Protoc. 2014;9(4):761–72.
20. Sidhu GK, Warzecha T, Pawlowski WP. Evolution of meiotic recombination genes in maize and teosinte. BMC Genomics. 2017;18(106):1–17.
21. Blary A, Gonzalo A, Eber F, Bérard A, Bergès H, Bessoltane N, et al. FANCM Limits Meiotic Crossovers in Brassica Crops. Front Plant Sci. 2018;9(368):1–13.
22. Zhang L, Tang D, Luo Q, Chen X, Wang H, Li Y, et al. Crossover formation during rice meiosis relies on interaction of OsMSH4 and OsMSH5. Genetics. 2014;198(4):1447–56.
23. Barakate A, Higgins JD, Vivera S, Stephens J, Perry RM, Ramsay L, et al. The Synaptonemal Complex Protein ZYP1 is required for imposition of meiotic crossovers in Barley. Plant Cell. 2014;26(2):729–40.

24. Liu Y, Schiff M, Marathe R, Dinesh-Kumar SP. *Tobacco Rar1*, *EDS1* and *NPR1/NIM1* like genes are required for N-mediated resistance to tobacco mosaic virus. Plant J. 2002;30(4):415–29.
25. Nimchuk Z, Marois E, Kjemtrup S, Leister RT, Katagiri F, Dangl JL. Eukaryotic fatty acylation drives plasma membrane targeting and enhances function of several type III effector proteins from *Pseudomonas syringae*. Cell. 2000;101(4):353–63.
26. Vaghchhipawala Z, Rojas CM, Senthil-Kumar M, Mysore and KS. Agroinoculation and Agroinfiltration: Simple tools for complex gene function analyses. In: A. P, editor. Plant reverse genetics Methods in Molecular Biology. Humana Press, Totowa, NJ; 2011. p. 65–76.
27. Peterson R, Slovin JP, Chen C. A simplified method for differential staining of aborted and non-aborted pollen grains. Int J Plant Biol. 2010;1(2):66–9.
28. Ross KJ, Fransz P, Jones GH. A light microscopic atlas of meiosis in *Arabidopsis thaliana*. Chromosom Res. 1996;4(7):507–16.

SUPPLEMENTARY MATERIALS

Additional supplementary Tables S3-S15 can be found online via: biorxiv.org/content/10.1101/459016v1.supplementary-material.

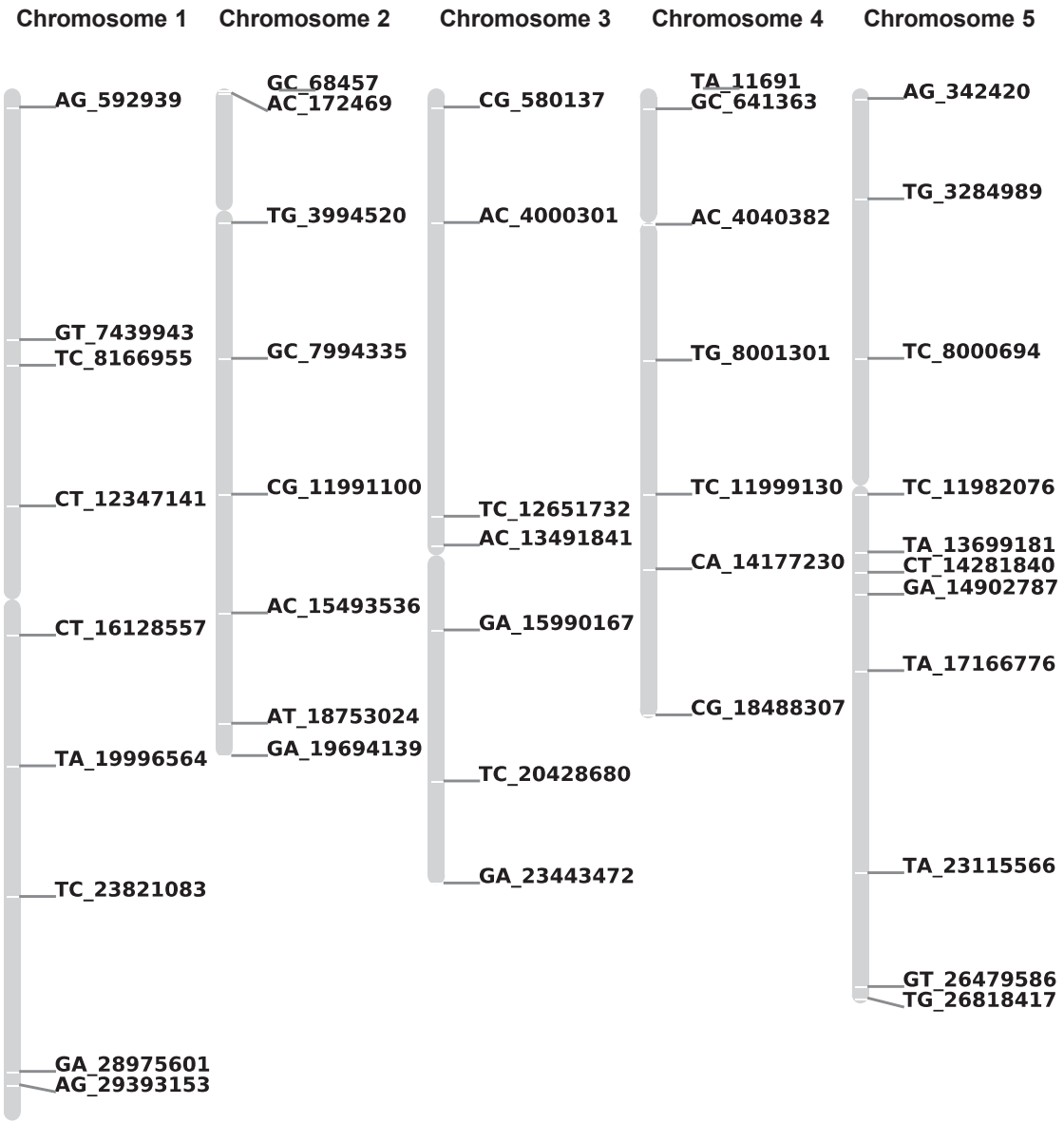


FIGURE S1 | Physical positions of markers used to genotype reverse breeding offspring. The names of used markers indicate the Col-0 allele, the L*er* allele and the bp position in the Col-0 reference genome.

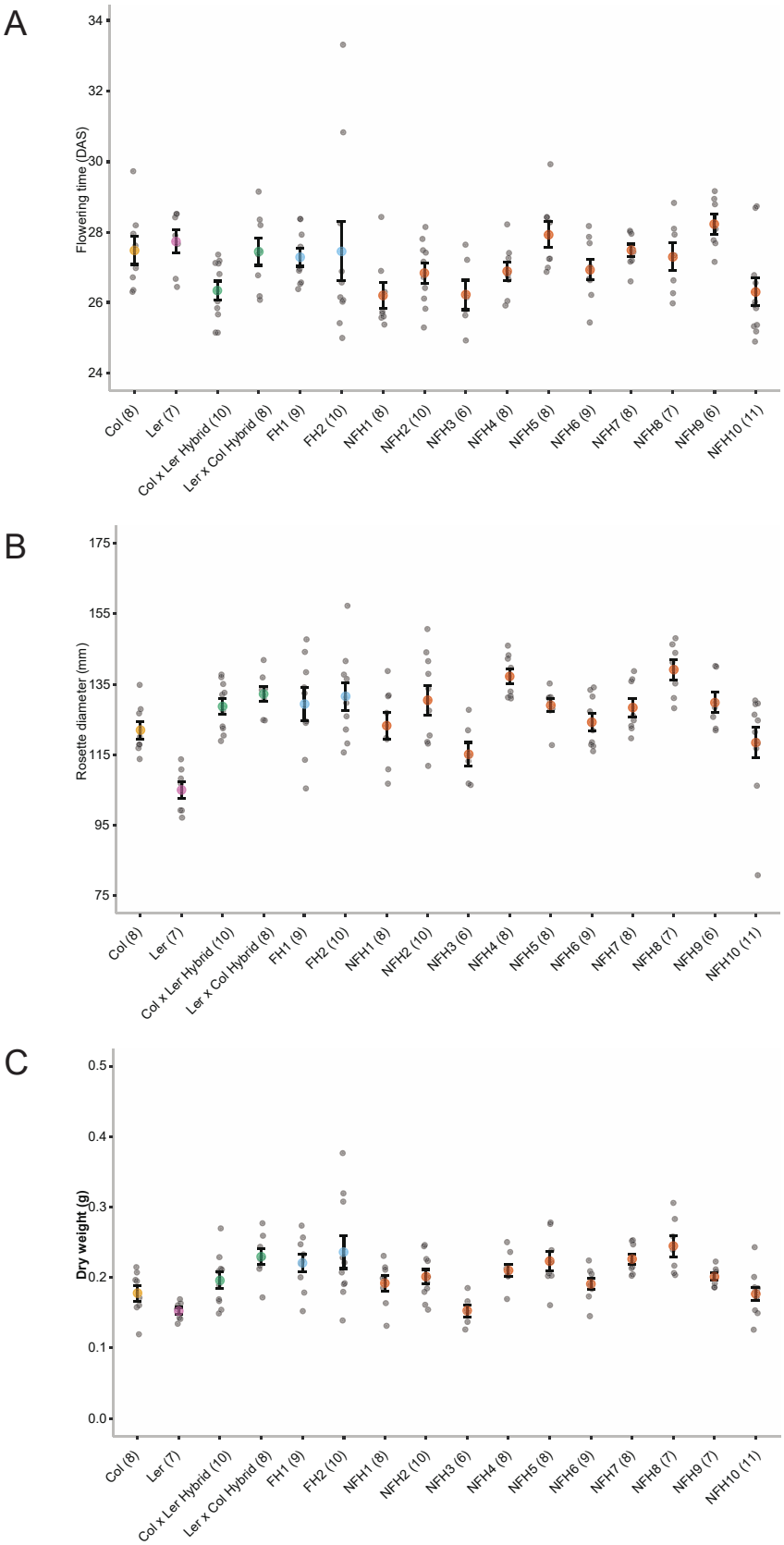


FIGURE S2 | The phenotypes of parental lines, full hybrids and partial hybrids for dry weight (A), flowering time (B) and rosette diameter (C). From left to right box-and-whisker plots re shown for the parental lines Col-0 (yellow) and L*er* (pink), Col-0 x L*er* reciprocal hybrids (green), full hybrids (FH, blue) and near-full hybrids (NFH, orange). FH and NFH genotypes are shown in the supplementary file.

TABLES S1-S6 | Expected number of DH_0 and DH_1 in DH populations with low crossover incidence. Tables indicate for different haploid chromosome numbers (1st row) the estimated numbers of crossovers per genome (CO/genome), the expected percentage of viable gametes and the expected percentage of DH_0 and DH_1 among DH offspring. The expected numbers of DH_0 and DH_1 are shown for the presence of 2 CO, 1 CO, 0.5 CO, 0.2, 0.1 and 0 COs per bivalent pair in tables 1 through 6 respectively. In many plants, like in Arabidopsis, the typical number of COs per bivalent pair is between 2 and 1 (i.e. tables 1 and 2). In the case of complete CO suppression, no DH_1 are recovered since no COs occur. For all calculations an exact (i.e. the indicated) number of COs per chromosome pair was assumed.

TABLE S1 | 2 COs per chromosome pair

n	CO / genome	Viable gametes (%)	DH_0 (%)	DH_1 (%)
1	2	100	25,0	50,0
2	4	100	6,25	25,0
3	6	100	1,56	9,38
4	8	100	0,39	3,13
5	10	100	0,098	0,98
6	12	100	0,024	0,29
7	14	100	0,0061	0,085
8	16	100	0,0015	0,024
9	18	100	0,00038	0,0069
10	20	100	0,000095	0,0019
11	22	100	0,000024	0,00052
12	24	100	0,0000060	0,00014

TABLE S2 | 1 CO per chromosome pair

n	CO / genome	Viable gametes (%)	DH_0 (%)	DH_1 (%)
1	1	100	50,0	50,0
2	2	100	25,0	50,0
3	3	100	12,5	37,5
4	4	100	6,25	25,0
5	5	100	3,13	15,6
6	6	100	1,56	9,38
7	7	100	0,78	5,47
8	8	100	0,39	3,13
9	9	100	0,20	1,76
10	10	100	0,10	0,98
11	11	100	0,049	0,54
12	12	100	0,024	0,29

TABLE S3 | 0.5 COs per chromosome pair				
n	**CO / genome**	**Viable gametes (%)**	**DH_0 (%)**	**DH_1 (%)**
1	0,5	75,0	75,0	25,0
2	1	50,0	56,3	37,5
3	1,5	37,5	42,2	42,2
4	2	25,0	31,6	42,2
5	2,5	18,8	23,7	39,6
6	3	12,5	17,8	35,6
7	3,5	9,38	13,3	31,1
8	4	6,25	10,0	26,7
9	4,5	4,69	7,51	22,5
10	5	3,13	5,63	18,8
11	5,5	2,34	4,22	15,5
12	6	1,56	3,17	12,7

TABLE S4 | 0.2 CO per chromosome pair				
n	**CO / genome**	**Viable gametes (%)**	**DH_0 (%)**	**DH_1 (%)**
1	0,2	60,0	90,0	10,0
2	0,4	35,0	81,0	18,0
3	0,6	20,0	72,9	24,3
4	0,8	11,3	65,6	29,1
5	1	6,25	59,0	32,8
6	1,2	3,75	53,1	35,4
7	1,4	2,19	47,8	37,2
8	1,6	1,25	43,0	38,2
9	1,8	0,70	38,7	38,7
10	2	0,39	34,9	38,7
11	2,2	0,23	31,4	38,3
12	2,4	0,14	28,2	37,6

TABLE S5 | 0.1 CO per chromosome pair				
n	**CO / genome**	**Viable gametes (%)**	**DH_0 (%)**	**DH_1 (%)**
1	0,1	55,0	95,0	5,00
2	0,2	30,0	90,3	9,50
3	0,3	16,3	85,7	13,5
4	0,4	8,75	81,5	17,1
5	0,5	4,69	77,4	20,4
6	0,6	2,50	73,5	23,2
7	0,7	1,33	69,8	25,7
8	0,8	0,70	66,3	27,9
9	0,9	0,37	63,0	29,9
10	1	0,20	59,9	31,5
11	1,1	0,11	56,9	32,9
12	1,2	0,06	54,0	34,1

TABLE S6 | 0 CO per chromosome pair				
n	**CO / genome**	**Viable gametes (%)**	**DH_0 (%)**	**DH_1 (%)**
1	0	50,0	100,0	0,00
2	0	25,0	100,0	0,00
3	0	12,5	100,0	0,00
4	0	6,25	100,0	0,00
5	0	3,13	100,0	0,00
6	0	1,56	100,0	0,00
7	0	0,78	100,0	0,00
8	0	0,39	100,0	0,00
9	0	0,20	100,0	0,00
10	0	0,10	100,0	0,00
11	0	0,049	100,0	0,00
12	0	0,024	100,0	0,00

Chapter 6

General Discussion

For unravelling genes and alleles underlying complex traits researchers have mostly adopted a reductionist view to study single loci and their interactions in a single (or few) genetic reference background(s) (*1*). While this is useful for the identification of the function of a gene, the impact a gene has can be dependent on the genetic background in which it is studied (*2, 3*). In this thesis phenotypic effects have been studied in homozygous *Arabidopsis thaliana* populations that can be obtained as a result of the availability of a genome elimination line for this species. This allowed a systematic mapping of quantitative traits in segregating populations. In some of the populations the genomic architecture was greatly reduced by a strong reduction or absence of crossover recombination. This reduction of complexity of the genomic architecture subsequently permitted the study of genetic interactions at the level of chromosomes.

In **chapter 2** a wild-type doubled haploid population was generated, such a population contains less crossovers as compared to the typical homozygous recombinant inbred lines (RILs) in which crossover events accumulate during inbreeding. In the subsequent chapters doubled haploids are used that are derived from an F_1 hybrid in which the crossover recombination frequency was largely or completely reduced. The stable transformation of an RNAi construct targeting DISRUPTED MEIOTIC cDNA (DMC1; **Chapters 3 & 4** (*4*)) or the transient silencing of MUTS HOMOLOGUE 5 (MSH5) using virus-induced-gene-silencing (**Chapter 5** (*5*)) allowed the production of homozygous genotypes with no or a few crossovers, respectively. The segregation of only non-recombinant chromosomes in such genotypes results in so-called chromosome substitution lines (CSLs).

It was investigated how the study of quantitative trait regulation could benefit from such novel doubled haploid resources. First, the use of haploid induction by a *GFP-tailswap* line (hereafter referred to as "*genome elimination*" line) was explored and it was demonstrated that a haploid generation can be used to the advantage of QTL mapping (**Chapter 2**). Second, it was presented how populations derived from the genome elimination line in combination with crossover suppression lead to variation at the phenotypic (**Chapter 3**) and molecular level (**Chapter 4**). Variation can be specifically attributed to chromosomal effects but also to inter-chromosomal effects caused by specific (combinations of) chromosome substitutions. Thirdly, it was shown that it is possible to develop more effective ways to generate non- and low-recombinant DH offspring (**Chapter 5**).

The current chapter highlights some of these topics in more detail and describes the potential implications for the scientific and breeding community. The genome elimination line was used throughout experiments in this thesis to generate doubled haploids. Because it has not been used extensively for the development of mapping

populations as of yet it is interesting to discuss more in-depth on the use of this genome elimination line. Next, both the advantages and disadvantages of CSLs are discussed with emphasis on the detection of epistasis. Although CSLs are advantageous for epistasis detection they can provide a versatile tool for studying other genetic concepts as well. Here several of these are outlined. Lastly, a discussion follows on the different leads that research on CSLs in *Arabidopsis* has provided for possible application in a plant breeding setting.

The use of the genome elimination line for DH production

Up to 2010 Arabidopsis was one of the major plant species for which no DH protocols were available (*6, 7*). Consequently, researchers in the Arabidopsis field have to "catch up" with crop researchers in exploring the use of DH research. Uniparental genome elimination in Arabidopsis by modification of CenH3 proteins has many different applications as was presented in a number of articles that followed its initial publication (*8-11*).

Both the fact that genome elimination takes place after fertilization and that it is uniparental means that haploid populations with new variation can be obtained from a single or few crosses using F_1 hybrid pollen (*8, 9*). For example, a single cross with a haploid inducer line can be enough to obtain numerous different recombinant haploids. This is in contrast to some of the current *in vitro* techniques where much effort is made to grow gametophyte tissues into mature plants (*7, 12-14*). This advantage makes the genome elimination technique suitable for the development of large segregating populations as was also shown in **chapter 2** of this thesis.

Besides the value for progressing genetic research in Arabidopsis, modification of CenH3 proteins was quickly suggested as a novel method for haploid induction in crops (*11*). During the past few years some of the initial ideas for breeding applications of uniparental genome elimination based on *cenh3*-modification have become reality. Where for maize and barley these applications are published in the public domain (*15, 16*), for other crops it seems that the private sector has taken the lead in making advancements by filling patents; for cucumber and melon (*17*), and tomato and rice (*18*). This signifies that *cenh3*-based uniparental genome elimination is emerging rapidly as a possibility for haploid induction in a number of crops (*12, 13, 19*).

Still, the efficiency of uniparental genome elimination and subsequent haploid induction has been very low for crop species such as maize and barley (*15, 16*). In crops it has been determined that specific genotypes can have different efficiencies for *in vitro* haploid induction (*7*). Therefore it is worth to reflect on this aspect with

the knowledge obtained from the DH-populations developed in this thesis. In **chapter 2** about 250 seeds were harvested from a relatively small number of crosses, enough to obtain a relatively large DH population. However, the efficiency of haploid production was exceptional for this specific cross (T540 x Ge-0 F_1 hybrid) as was later recognised when other F_1 crosses were tested. For instance the Col x L*er* F_1 hybrid was much less efficient in generating haploids, even though a similar number of crosses were made (*20*).

The high haploidisation efficiency (although not explicitly quantified) for the T540 x Ge-0 F_1 hybrid might be based on a fortunate chance. However, the use of additional crosses of F_1 hybrids derived from either the T540 or Ge-0 accessions seems to indicate that especially the T540 genotype is favouring haploid induction and thus a genotype effect might be causal for higher efficiencies (R. Botet, unpublished data). Obtaining many haploid lines prior to the development of the T540 x Ge-0 DH population indicates a high efficiency in both genome elimination and seed set. In case a genetic factor is involved this would most likely make a gamete more competent to cross-fertilise the genome elimination line and to zygote survival during the initial phases of genome elimination. Supporting such a hypothesis the causal allele would be overrepresented in the subsequent DH population. In the T540 x Ge-0 DH population there is a relative large allele frequency distortion on chromosome 1 in favour of the T540 genotype (**Chapter 2 Fig. S2**). This region would therefore certainly be of interest to investigate further in the context of obtaining higher efficiencies of DH induction.

Taken together, *cenh3*-based genome elimination is slowly taking centre stage for haploid induction in crops. The advantages over recombinant inbred lines (RILs) and current *in vitro* DH methods for obtaining homozygous lines is being appreciated in both the scientific and breeding communities. Still, the efficiency in crops has been very low so far and it is likely that further investigations to achieve higher efficiencies and translate the technique to other crop species will follow the current successes. Especially in combination with the reduced recombination rates proposed in **chapter 5** this would make reverse breeding a possibility for higher chromosome species.

One of the most recent and impressive new additions to *cenh3*-based haploid induction has been the combination of genome elimination and CRISPR-Cas9 technology (*21*). In an enlightening paper it was shown how genome elimination coupled with CRISPR-Cas9 constructs can immediately edit the genome of the haploid offspring. This was not only shown in *Arabidopsis* but also in maize and wheat. In their experiments with Arabidopsis the activation of the CRISPR-Cas9 before genome elimination was sufficient to edit the non-eliminated genome conveying directly a mutant phenotype to the growing haploid plant. This demonstration of the

application of both CRISPR-Cas9 and haploid induction via CenH3 modifications to plant breeding is inspiring and shows great promise for future research.

Considerations of using CSLs for genetic mapping

The identification of genotype-phenotype relationships is one of the key drivers behind quantitative genetic research and evolutionary biology. QTL analysis has provided procedures to study trait associations (*22*) and with new methodologies QTL are detected with increasingly better resolution. Still, major challenges remain. QTLs commonly have large confidence intervals since only for large populations of more than 1000 individuals confidence intervals become small enough when QTLs have a large explained variation. Additionally, detected QTLs are not always confirmed when their effect is studied in a different genetic background (QTL x genetic background) (*23*). These difficulties are accompanied by multiple QTLs on the same chromosome for quantitative traits which can lead to problems in detecting any of them (*24*) and epistasis which further complicates identification of causal loci because their effect on a trait value may depend on other loci in the genome.

One of the first descriptions of epistasis by Bateson handled about the atypical frequencies of flower colours in F_2 offspring of peas that showcased the dependency on the presence of at least two complementary genes in a single pathway (*25*). This is an obvious example how detection of epistasis can aid in resolving specific biochemical pathways and how genetic interactions can increase trait variation. From previous research it is clear that epistasis affects many, if not all, phenotypic levels including molecular and developmental traits (*26, 27*). Two cases of background dependent genetic effects in Arabidopsis show that developmental phenotypes such as circadian rhythm and flowering time are clearly regulated by epistatic interactions (*28, 29*).

The history of CSLs dates back more than half a century (*30*), still **chapter 3** describes the first panel of all possible chromosome combinations in a segregating population in any species (*4*). This unique panel has allowed the investigation of chromosomal interactions and identified that epistatic effects can be as large as main effect QTLs. Now that many QTLs for common traits have been identified it is time to look beyond the low-hanging fruits and identify the interaction effects that may cause the exhaustive variation that can be present within a species.

In **chapter 3** the analysis of this complete CSL panel including all genotypes shows the possibility to determine which specific combination of chromosomes is causal for epistatic interactions (*4*). Clarifying the exact magnitude of epistatic variance and its molecular architecture is of importance for predicting the manifestation of

phenotypic traits from genetic information. Finemapping still needs to be done so it will take an additional effort to understand the molecular basis of epistasis by positional mapping of higher-order interactions detected in CSLs as will be discussed later in this chapter. Still it needs to be considered what the direct benefit is of this knowledge of epistasis, and what can be done with such knowledge.

The gain of understanding epistatic interactions

The ultimate goal in plant breeding programmes is to optimize the genetic gain, i.e. the increase in trait performance per generation. Genetic gain is measured with the breeders equation, (*31, 32*). Here genetic gain per breeding cycle is the product of the additive genetic variance , the selection intensity , and the selection accuracy divided by the number of cycles per year (). It is clear from this equation that the explained genetic variance is a major aspect for the genetic gain that can be achieved. With the increased knowledge of epistatic interactions more genetic variance is explained for phenotypic traits and hence the information on epistasis ultimately leads to higher genetic gain.

An specific example where the knowledge of epistatic interaction can be used to the benefit of optimizing genetic gain is in genomic selection. With genomic selection traits are regressed against all genetic markers for a training population to subsequently predict performance of offspring based on only their genotype (*33*). Such models rely on the additive effect of single markers to estimate genomic estimated breeding values (GEBVs). However, these models could benefit from knowledge on epistatic complexes (*34, 35*). Especially in the case of (self-fertilizing) species or quantitative traits with genetic architectures that consist mainly of interactions between small effect QTLs the addition of epistatic information can improve the predictive ability of such models (*36-38*). The current predictive models do not explicitly model interactions but they are expected to benefit from explicit knowledge of epistatic components (*33, 39*). Examples of increased predictability when including epistasis was provided in **chapters 3 & 4**, where models including epistasis explained almost 10% more variation for morphological or even 25% for molecular traits than models without epistasis.

Also the suggested use of optimal haploid values (OHVs) can benefit from the addition of epistatic information (*40*). The calculation of OHVs is largely similar to that of GEBVs but instead of describing the estimated breeding value of a certain genotype the OHV estimates the best possible DH offspring that may result from a specific cross. An important aspect in which OHVs differ from GEBVs is that OHVs allow combinations of haplotypes (in contrast to single markers) that optimize genetic gain. In that respect, the identification of specific inter-chromosomal interactions that

have a desired effect on a phenotype can easily be included as a single haplotype required for optimal gain.

More specifically, it would be interesting to incorporate knowledge obtained from the CSLs and estimate OHVs and compare those with DH offspring for confirmation. For instance, for multiple traits CSLs with both chromosomes 2 and 5 from the Col parent show much larger between line variation in contrast to CSLs where either or both of these chromosomes are descending from the L*er* parent. This was most evidently demonstrated in **chapter 4** in which chromosome 2 and 5 are shown to have a large impact on proteome variation. Such knowledge is valuable and might improve the predictive power of models that are currently based on only additive effects.

A comprehensive view on additive vs. epistatic effects

In **Chapter 3** it was described how CSLs can be used for the detection of main effects and epistatic interactions. It is interesting to once more discuss the outcomes of the analyses in that chapter because there are more ways in which epistasis could be studied in a CSL context. A main starting observation in **chapter 3** shows how sCSLs (single introgressed chromosomes in a parental background) are useful for the detection of main effects. But immediately it was emphasized how the use of a reciprocal set of sCSLs could determine whether a QTC is also detected in a different genetic background. The last few years such background effects are getting more and more attention (*2, 3*).

An example of such an interaction effect is the detection of QTCs for main stem length in the Col sCSL set but not in the L*er* set (**Chapter 3: Fig. 2A-B**). With all possible sCSLs the effect of the substitution of a single chromosome can be tested in two backgrounds (e.g. AAAAA vs. BAAAA and ABBBB vs. BBBBB for comparing chromosome 1, where A and B stand for different genotypes at chromosome (or locus) 1 through 5, respectively). However, in a complete CSL panel many more of such comparisons can be performed. To be exact, sixteen complementary pairs of two near-isogenic genotypes differing only in the genotype at chromosome 1 can be compared. Of course, such comparisons can be extended to genotypic variation at all five chromosomes forming a total of eighty independent tests of single chromosome substitution effects.

If all those separate comparisons of a single chromosome substitution were analysed one might expect that a main effect QTC identified in one comparison is detected in each of the sixteen comparisons. However, it is clear that this is not always the case even when the overall analysis indicates a highly significant main effect which

is a realistic scenario in RIL mapping. This also implies that when a chromosome substitution is analysed in the wrong background a false negative conclusion will be drawn. However, including the analysis of different combinations will reveal that chromosome substitution effects might be context (or genetic background) dependent. Surprisingly, genetic interaction effects are often ignored in biological studies (*41*) and unilateral comparisons are typical for the reductionist view on science (e.g. in reverse genetic studies). The observation that the detection of QTCs often depends on the genetic background is therefore a first step towards the unravelling of epistatic interactions as there are clear dependencies of QTCs on the genetic background.

The observation that main-effects are background dependent also applies to the detection of two-way or even higher-order interactions For the detection of two-way interactions four two-locus haplotypes (AA; AB; BA and BB) need to be compared instead of just two. Adopting a similar approach as was done for single chromosome effects and assuming no higher-order interactions are involved, in a complete CSL panel each haplotype consists of eight different combinations (i.e. backgrounds). Since ten different two-way inter-chromosomal effects can occur in the Arabidopsis genome consisting of five chromosomes, a total of eighty haplotype comparisons can be made and investigated. In case of three-way interactions the number of haplotypes to be compared increases two-fold, while the number of lines representing a single haplotype would reduce two-fold. Consequently, since also ten possible three-way interactions can occur the number of comparisons remains the same.

The aim of this exercise is to illustrate that taking an analytical approach on main effects and epistasis that does not consider the population as a whole but rather a limited set of CSLs can reveal specific epistatic interaction effects. However applying an analysis of epistasis on a complete CSL panel allows to sketch a good general view on genetic interactions at play. Therefore it is important to keep in mind that such an approach does present a limited view of the total number of interactions that are present as it does not exclude the possibility of significant interactions occurring only in a specific genetic background. This indicates that the genome is much more complex than can be described by a simple regression model and it shows the dependency of phenotypes on a large number of genetic and environmental factors. Instead of a simple regression model, it might be worthwhile to incorporate a strategy based on mixed models to describe the complexity of the genome. With a complete CSL panel, the tool has been provided for a step-by-step approach towards decrypting this genomic complexity.

Finemapping inter-chromosomal epistasis

It is important to note that while there are many studies identifying QTLs, there is a general lack of QTL cloning studies (*42, 43*). This thesis as well can be accused of providing only new leads and candidate genes for further exploration. Together with the CSLs in **chapter 3** the interesting possibility of finemapping detected higher-order interactions to a finer scale is introduced. By using the doubled haploid near-isogenic lines (DH-NIL) main effects were coarsely mapped. Similarly, the CSL genotypes allow the finemapping of genetic interaction effects. Below, an example is given for the finemapping of the three-way interaction explaining main stem length identified in **chapter 3**.

First, it needs to be noted that a significant interaction as identified by the models in **chapter 3** are presumed to be independent of the composition of the remainder of the genome. However, the estimated interaction effect size is based on multiple genotypes that can affect the size of the residual genetic variation. For instance, for the significant interaction Chr1^{Ler}/Chr2^{Col}/Chr5^{Ler} the genotype of chromosomes 3 and 4 displayed no significant effect on main stem length. This said, the genotype LC**LL**L was on average 17% taller than the LC**CC**L genotype (139.5 vs. 118.9 mm, respectively; where C and L stand for a Col or L*er* origin of the chromosome) (**Fig. 1A**).

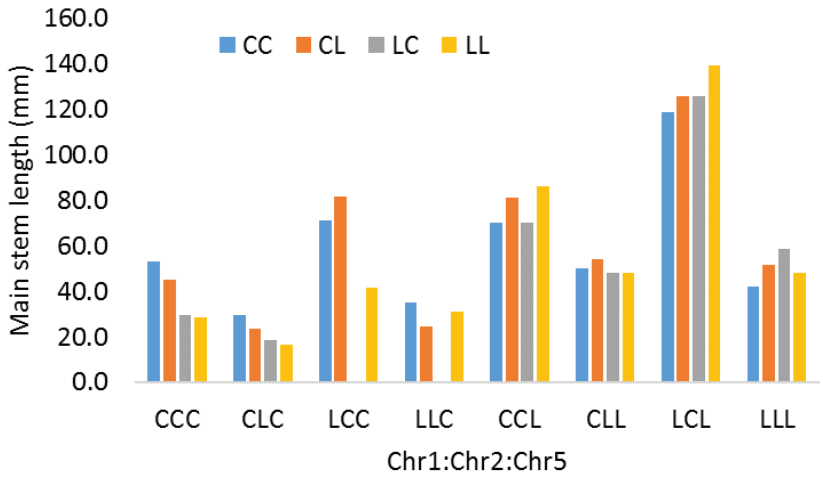


FIGURE 1 | Finemapping three-way epistasis for main stem length. The three-way epistatic interaction for Chr1/Chr2/Chr5 as it was identified for main stem length in chapter 3 is represented by each of the eight haplotype classes on the x-axis. The labels refer to the chromosomal composition at chromosome 1, 2 and 5, respectively where C indicates a Col chromosome and L a L*er* chromosome. The bars depict the individual best linear unbiased predictor (BLUP) value for each of the genotypes and the colours indicate the different combinations of genotypes at chromosome 3 and 4.

Therefore, before continuing to fine-map the interaction the genotype with the largest contrast with other genotypes, in this case the genotype with the highest mean trait value, should be identified. Subsequently, to isolate each single component of the higher-order interaction this genotype needs to be crossed to the three CSLs that only differ for one of the chromosomes that take part in the three-way interaction. Here another advantage of a complete CSL panel is pointed out: because all possible genotypes are present in such a panel specific crosses can be performed such that F_1 hybrid offspring is created in which only the targeted chromosome is heterozygous. Theoretically it is even conceivable that in order to fine-map an interaction locus on a specific chromosome one may consider choosing the background such that the effect of that specific chromosome is most contrasting.

In the case of the $Chr1^{Ler}/Chr2^{Col}/Chr5^{Ler}$ three-way interaction this leads to three different partial hybrid F_1 lines that contain only a single heterozygous chromosome (**Fig. 2A**). Upon self-fertilization or when crossed to the haploid inducer line each F_1 yields an offspring population segregating for a single chromosome while all other chromosomes are homozygous for either parental genotype (**Fig. 2B**). Phenotypic screening of such offspring populations will be informative for the position of the QTL.

Of the three suggested populations only the one segregating for chromosome 2 was developed and investigated and it seems highly likely that ERECTA is the interacting genetic factor located on chromosome 2 (*44*). The other two populations still need to be created but could relatively quick determine the coarse position of the other two interacting loci which might lead to candidate gene selection. Considering the

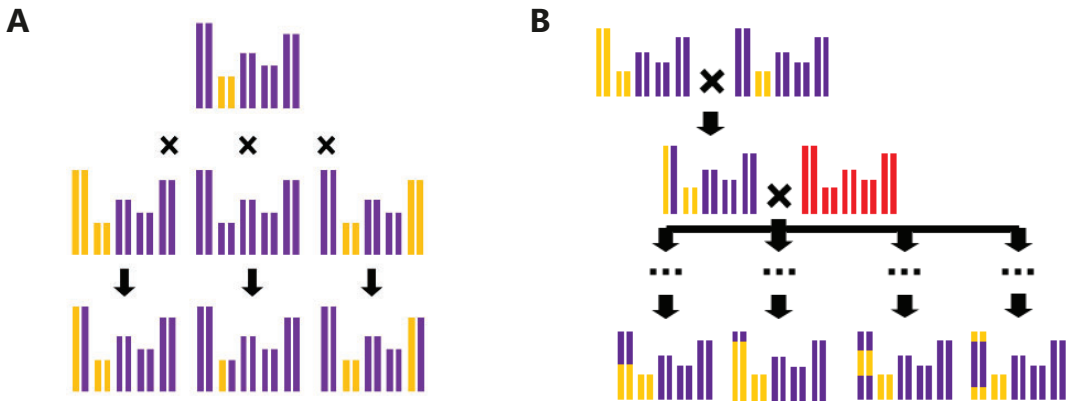


FIGURE 2 | Crossing scheme designs for fine-mapping the three loci underlying the three-way epistatic interactions. A) To ensure only a single chromosome will segregate in the offspring, while fixing the genotype of the two other chromosomes contributing to the epistatic effect, the genotype will be crossed to three lines with a contrasting genotype at only one of the three chromosomes. B) The F1 hybrids obtained from these crosses can be used as the paternal genotype in a cross with the genome elimination line (depicted in red) to obtain DH-NILs for positional mapping.

high number of genes ERECTA is known to interact with it is tempting to speculate which candidates might be located on chromosome 1 and 5. However, validation by functional analyses such as through up- or down-regulating gene expression or complementation studies should be performed before any confident claims can be made. It is also possible that not one but multiple loci reside on a specific chromosome that together contribute to the epistatic effect of the chromosome. Also this could be resolved during fine-mapping.

Reduced crossover recombination does not lead to less phenotypic variation

Meiotic crossover recombination is responsible for the admixture of genetic material before transfer to offspring which is often considered as one of the driving forces of phenotypic variation (*45*). In addition, crossovers are essential for viable gamete production as it is the mechanism by which homologs become connected by (at least one) chiasma to ensure proper disjunction to both daughter cells. Without a chiasma homologs segregate randomly during the first meiotic division which leads to high levels of aneuploidy and lethality among gametes. Therefore, a general requirement of at least a single crossover per chromosome pair is necessary for the production of viable gametes in most species. Still, as is shown in **chapter 3**, *Arabidopsis* is capable to produce viable gametes (and hence offspring) at low frequency in the absence of crossovers. This offers the unique opportunity of testing the impact of crossover recombination on phenotypic variation.

In this thesis, there have been two comparisons of two population types that differ in terms of crossover frequency: In chapter 2 a DH population was contrasted with an F2 population (homozygous lines with few crossovers vs. partially heterozygous plants with crossovers on both homologs) and in chapter 3 the comparison of CSLs (that show chromosome segregation but not crossover recombination) with RILs (that show both chromosome segregation and crossover recombination) was performed. This allows one to ask and explore what the effect of crossover recombination is on plant phenotypes.

A principal component analysis (PCA; **Fig. 3**) of segregating populations differing in the presence (CSL vs RIL) or frequency (DH vs F_2) of crossovers revealed that a reduction in crossover frequency does not lead to a decrease in phenotypic variation even though the genetic architecture of CSLs is reduced respectively to the populations to which these are compared. In fact, the comparison of the CSLs and RILs suggests that the phenotypic variation of the CSLs is actually larger than that of the RILs. It is interesting to speculate on possible causes of this dependency of phenotypic variation on genetic variation caused by crossover

recombination and/or chromosome segregation (for the current species, parental cross and phenotypes).

One explanation might be that crossovers cause disruption of intra-chromosomal epistatic interactions as extreme phenotypes are mostly observed for CSLs. Compared to outcrossing species, epistasis occurs more frequently in facultative inbreeding species like Arabidopsis (*46*). Self-fertilization allows the co-evolution and fixation of additive-by-additive epistatic loci over multiple generations and thus a relatively high epistatic variance can be expected compared to additive variance for selfing species (*47*). Outcrossing Arabidopsis to generate genetic variation in a mapping population might consequently disrupt epistatic interactions and reduce the phenotypic variance. Contrary in CSLs, in which intra-chromosomal epistasis remains preserved and the chromosomes contain allele combinations that arose in wild accessions (either through adaptation, drift, or otherwise) are the single blocks that drive phenotypes towards the extremes of the distributions. This larger phenotypic variation within the CSL panel can be beneficial for selection breeding and QTL mapping and again advocates for the use of CSLs for the initial detection of epistatic interactions.

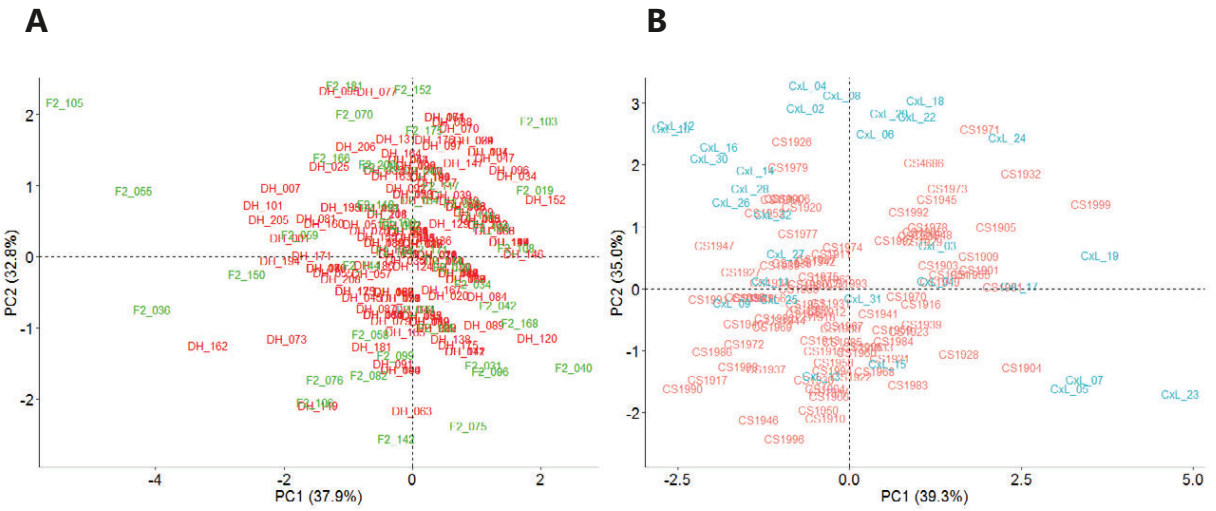


FIGURE 3 | Principal component analyses (PCA) of the different populations used in chapters 2 and 3. A) The PCA based on three phenotypes (main stem length, number of branches from rosette and number of branches from main stem) measured in both the F_2 (green) and DH (red) populations derived from the Ge-0 x T540 F_1 hybrid. B) The PCA plot based on five phenotypes (flowering time, main stem length, cauline leaves, rosette diameter and dry weight) measured in both the CSL (blue) and RIL (pink) population derived from the Col x L*er* F_1 hybrid.

OPPORTUNITIES PROVIDED BY THE FLEXIBLE USE OF CSLS

CSL panels allow the systematic dissection of heterosis

As was shown in **chapter 3**, with their small population size and simple genetic architecture CSLs provide the opportunity to separately explore the impact of additive and epistatic effects (non-allelic interactions). The study of heterosis (or hybrid vigour) with such a population was not given much specific attention so far. A trait is considered heterotic when the offspring performs superior either to the average phenotype of its two parents (mid-parent-heterosis; MPH) or to the best performing parent (best-parent-heterosis; BPH) (*1, 48*). CSLs were suggested as tools to estimate the contribution of epistasis and heterozygosity separately for heterotic traits in hybrid F_1 (*49, 50*).

While heterosis is often considered in heterozygous genotypes, heterozygosity is not per se required for heterosis. Recombination of genetic variants in crossing progeny can be enough to outperform the genetic constitution of founder lines. This is most clearly illustrated by the phenotypic transgression of crossing offspring. For instance DH offspring can outperform their F_1 hybrid parent or continuous inbreeding of heterozygous genotypes can lead to surpassing both parental as well as F_1 hybrid trait values referred to as hybrid mimicry (*51*). Similarly, in **chapter 5** the near-full hybrids demonstrate that different levels of heterozygosity result in phenotypically comparable hybrids. Currently there is still no consensus on how large the contribution of epistasis and heterozygosity is on the regulation of heterosis (*48, 52*) although it seems clear that heterosis is specifically dependent on the species and the investigated trait.

The systematic decomposition of the genome to detect epistasis as was performed in **chapter 3** can work equally well to disaggregate dominance, additive x dominance and dominance x dominance effects in a panel of intercrossed CSLs. For instance, the CSLs can be intercrossed in a full diallel design in which all 32 CSLs are used as parental genotypes that are crossed in every pairwise combination to generate F_1 hybrids in which all chromosomes are either a homo-homozygous for one of the founder genotypes (AA or BB) or heterozygous (AB) (**Fig. 4A**).

An independent analysis of epistasis for only homozygous genotypes (**chapter 3**) followed by the subsequent analysis of heterozygous genotypes can disentangle the proportion of epistasis and dominance effects to heterosis. A full diallel cross would involve a total of 32 x 32 (or 2^5 x 2^5) crosses to obtain 1,024 offspring hybrids. While such a number of crosses might be manageable when dedicated efforts are made smaller but still informative panels can be created by optimizing the design of such crosses.

When following a full diallel design many crosses result in duplicate genotypes. Although duplicate genotypes might be useful for certain purposes as will be discussed later, for many applications a reduction in population size is desirable. Considering three different allelic combinations per chromosome and five chromosomes for Arabidopsis, 243 (3^5) unique combinations of homo- and heterozygous chromosome combinations can be obtained (**Fig. 4B**). A panel of all of these 243 genotypes would be useful for investigating both heterozygous and epistatic effects simultaneously.

More practical designs for higher chromosome species could also be considered for a more applied breeding situation. For instance it could be *a priori* decided to have one or two fixed chromosomes containing desired QTLs in a population (**Fig. 4C**). This could already reduce the number of CSLs to be considered considerably. For two fixed chromosomes only those backcrosses have to be performed that result in genotypes with the two desired chromosomes homozygously fixed, which would result in only 64 genotypes (2^3 x 2^3) for Arabidopsis. Such a population would allow the evaluation of interactions with the fixed QTLs and the heterosis of the other chromosomes.

Identical hybrids: an informative study on the concept of heterosis

Heterosis of an offspring genotype has been credited to many other genetic concepts besides the aforementioned effects of heterozygosity and epistasis (*48, 53*). For instance parent-of-origin and ploidy effects can play a role as well (*54-56*). Again CSLs could be useful research tools for investigating contributions of such phenomena. An informative subset of hybrid genotypes obtained from intercrossing CSLs that could be used for studying these genetic concepts are the so-called identical hybrids. Identical hybrids are obtained by crossing different pairs of genetic complementary CSLs, which results in genetically identical offspring (**Fig. 4D**). In **chapter 5** one such cross was made to recreate the original wild-type hybrid and check whether such an identical F_1 resulted in a phenotypically similar plant as the wild-type hybrid.

Usually the creation of identical hybrids is limited to reciprocal crosses of two parents (P_1 x P_2 and P_2 x P_1). This limits the analysis of reciprocal hybrids to checking whether the mating direction of the cross is important or not. In a complete CSL panel of Arabidopsis a total of sixteen different complementary pairs are present that when crossed each result in genetically identical hybrid F_1 offspring. In case those complementing pairs are crossed reciprocally identical hybrids obtained from 32 genotypically different parent pairs can aid in the detection of heterosis and parent-of-origin effects.

It would also be interesting to investigate in such a panel of identical hybrids how their performance compares to their respective mid-parent values (MPVs). This could resolve the question whether the classification of heterosis is dependent on the MPV or if heterosis is limited by the genotype of the hybrid itself. Based on the current best linear unbiased predictors (BLUPs) of the CSL panel it is clear that MPVs for the different identical hybrids will not be identical (**Fig. 5**). Such an experiment could

FIGURE 4 | Schematic overview of a diallel design with different subsets. For clarity a diallel design for only three chromosomes is considered. On the top and left outside borders are the CSL genotypes coloured either red (AA), or blue (BB). The genotypes inside the boxes represent the F_1 hybrid offspring when crossing the corresponding parental lines located on the outside borders. To enhance visual inspection heterozygous chromosomes (AB) are coloured green. A) The full diallel design. Note how the F_1 hybrids in the box contain many duplicate genotypes, including the completely homozygous CSL genotypes on the diagonal top left – bottom right. B) The unique genotypes within the diallel. C) The genotypes when considering a single fixed chromosome (here chromosome 1). D) The diagonal of the diallel contains genetically identical F_1 hybrids.

have important implications for the general concept of heterosis. If the different identical hybrids show a constant phenotype it would indicate that the genotype of the hybrid is the main driver behind heterosis. This would also imply that the quantification of heterosis differs depending on the specific parental cross under examination considering the difference in MPVs.

							Constant hybrid		Constant heterosis	
Chr1	Chr2	Chr3	Chr4	Chr5	MSL (mm)	MPV (mm)	MSL (mm)	Heterosis (%)	MSL (mm)	Heterosis (%)
Ler	Col	Ler	Ler	Ler	139.5	84.5	75.0	-11.2	105.6	25.0
Col	Ler	Col	Col	Col	29.5					
Ler	Col	Ler	Col	Ler	125.5	74.4	75.0	0.7	93.1	25.0
Col	Ler	Col	Ler	Col	23.4					
Ler	Col	Col	Ler	Ler	125.9	72.3	75.0	3.7	90.4	25.0
Col	Ler	Ler	Col	Col	18.8					
Ler	Col	Col	Col	Ler	118.9	67.7	75.0	10.8	84.6	25.0
Col	Ler	Ler	Ler	Col	16.4					
Ler	Col	Col	Ler	Col	81.6	64.9	75.0	15.6	81.1	25.0
Col	Ler	Ler	Col	Ler	48.2					
Col	Col	Ler	Ler	Ler	85.9	60.5	75.0	24.0	75.6	25.0
Ler	Ler	Col	Col	Col	35.0					
Ler	Col	Col	Col	Col	71.3	59.7	75.0	25.6	74.6	25.0
Col	Ler	Ler	Ler	Ler	48.1					
Ler	Ler	Ler	Col	Ler	58.5	51.8	75.0	44.8	64.8	25.0
Col	Col	Col	Ler	Col	45.1					
Col	Col	Col	Col	Ler	70.1	50.7	75.0	48.1	63.3	25.0
Ler	Ler	Ler	Ler	Col	31.2					
Col	Col	Col	Col	Col	53.1	50.5	75.0	48.5	63.1	25.0
Ler	Ler	Ler	Ler	Ler	48.0					
Col	Col	Ler	Col	Ler	70.1	47.4	75.0	58.2	59.3	25.0
Ler	Ler	Col	Ler	Col	24.7					
Col	Ler	Col	Col	Ler	49.9	45.8	75.0	63.9	57.2	25.0
Ler	Col	Ler	Ler	Col	41.6					
Ler	Ler	Col	Ler	Ler	51.6	40.7	75.0	84.4	50.8	25.0
Col	Col	Ler	Col	Col	29.8					
Ler	Ler	Col	Col	Ler	41.9	35.4	75.0	112.1	44.2	25.0
Col	Col	Ler	Ler	Col	28.8					

FIGURE 5 | Numerical considerations of the underlying mechanisms of heterosis. The first five columns indicate the composition of the genotype. Each combination of two rows shows the complementing pair of genotypes that form a full hybrid genotype upon crossing. The phenotypic values (BLUPs for main stem length (MSL)) for the two CSLs with their respective mid-parent value (MPV) are given in the next two columns. The last four columns present two hypotheses: 1) Constant hybrid performance and 2) constant heterosis level. The constant hybrid hypothesis dictates that the phenotype of the identical hybrids is equal due to their identical genomes but due to different MPVs the percentage of heterosis will differ. The constant heterosis hypothesis assumes an equal relative additive effect of heterosis on top of each MPV, with the consequence that genetically identical genotypes perform differently from one another.

In a second scenario the level of heterosis is pervasive and constant suggesting that the specific genotype of the complementing parents determines the hybrid performance. This would imply that the genotype of the hybrid itself although identical, is of less influence on the hybrids performance. In this second scenario it is likely that parent-of-origin effects have a major influence. With multiple identical hybrids of different parental origins the possibility exists to identify the specific parental chromosomes that show such a transgenerational effect. The sets of identical hybrids created by intercrossing CSLs can thus be extremely helpful to explore and decompose heterosis, mating direction and also parent-of-origin effects.

Detection of ploidy effects with chromosome substitution lines

Yet another application discussed here is the investigation of the effects of ploidy on the performance of F_1 hybrids. A recent study separated the effect of ploidy level and heterozygosity on heterosis by comparing rosette growth speed among diploid, tetraploid and reciprocal triploid (2n = 3x) lines for ten *Arabidopsis* accessions (*54*). In this study especially Col and L*er* responded differently to the increase in ploidy level, which led to consider how to exploit the available complete CSL panel for this purpose.

Research on ploidy differences in Arabidopsis generally only concerns the comparison between isogenic and/or hybrid lines of a biparental cross. So far, a dedicated mapping resource to dissect the effects of ploidy has not been used. This reflects the difficulty of developing a homozygous polyploid mapping population. The development of such populations is time consuming and substantial investments are needed for generating and validating the high number of genotypes that informative populations are typically composed of. In addition, the development of a polyploid mapping resource adds an extra layer of complexity, which is generally not wanted.

Again the small size of the CSL panel makes it much more feasible to create a homozygous autotetraploid population of CSL genotypes. Such a panel limited to the sCSLs in which only a single chromosome is substituted should already be informative to identify chromosomal effects (albeit in a single genotypic background). The comparison of such a population with its diploid complement allows for the detection of chromosome dependent ploidy effects. It would be interesting to see if such a panel indeed can detect the so called genotype-by-ploidy (GxP) interactions described in **chapter 2**.

Small polyploid CSL panels can be extended by including mono- and triploid genotypes or by complementing the sCSL panel with all possible polyploid CSLs. Monoploids can be created by crossing diploids to the genome elimination line and

triploids are obtained from crosses between di- and tetraploids. Evaluating a panel of lines with different ploidy levels and genotypes can be highly informative for research on polyploidy. Indeed, where the focus of this thesis has been on the use of CSLs in a diploid plant species many crops are not diploid. For these, polyploid panels can be used to create different parental chromosome dosages (*e.g.* AAA; AAB; ABB and BBB for parental genotypes A and B in a triploid genome) and as such genome complexity can be studied one level at a time.

Additional applications of CSLs

Besides studying epistasis and heterosis other possibilities for studying genetic concepts with CSLs spring to mind, not the least the use of CSLs in analysing genotype-by-environment (GxE) and QTL-by-environment (QxE) effects (*57, 58*). GxE effects can be substantially large as was demonstrated for flowering time in Arabidopsis (*59*) and for the DH population in **chapter 2**. The study of GxE effects focusses on variation that occurs when identical genotypes are grown at different locations or in different conditions. Such a study can easily be performed with homozygous CSL panels which are small enough to be replicated to a high degree in multiple environments.

The use of multiple parents to increase the allelic variation is another trend that can create added complexity to the genetic architecture of a mapping population (*43, 60-63*). Of course similar approaches to increase allelic (chromosomal) variation can be applied to a CSL population. One can imagine how the use of additional chromosomal variation in a population can lead to extra phenotypic variation that can be taken advantage of for the investigation of additional complex interactions.

CONCLUDING REMARKS

The practical use of CSLs in genetic mapping since the second half of the last century has gone through some high and low times (*30*) but since the beginning of this century CSLs have gained attention in different model species (*64-66*). From the first moment it was clear that these panels limited to genotypes with substitution of a single chromosome (sCSLs) in a single recurrent background could contribute to the detection of epistasis in complex traits (*67*). While small populations like these are informative of the existence of epistasis, the localisation of epistatic loci can only be investigated by substitution of multiple chromosomes simultaneously ((*68*) and **chapter 3 & 4** of this thesis). This suggests that other possible designs of substitution panels in which sCSLs are intercrossed could be informative as is also suggested in **chapter 5**. Still, such panels would suffer from the possibility that the detected effects are conditional to a specific recurrent genetic background (*2, 3*). My endeavours clearly indicate that the best solution to obtaining a complete picture of the genetic architecture controlling quantitative traits is by obtaining a complete CSL panel with all possible combinations of substitutions as is shown in **chapter 3** (*4*). However, for practicality reasons in high chromosome number species, it needs to be considered how to apply CSLs and if smaller panels might be informative enough for the purpose of genetic mapping. These considerations should include the type of genetic effects one wants to identify and the genetics of the considered trait. Most importantly, CSLs are tools that provide additional information on the genetic architecture of quantitative traits.

REFERENCES

1. M. Lynch, B. Walsh, *Genetics and analysis of quantitative traits*. (Sinauer Sunderland, MA, 1998), vol. 1.
2. C. H. Chandler, S. Chari, I. Dworkin, Does your gene need a background check? How genetic background impacts the analysis of mutations, genes, and evolution. *Trends in Genetics* 29, 358-366 (2013).
3. J. Hou, G. Tan, G. R. Fink, B. J. Andrews, C. Boone, Complex modifier landscape underlying genetic background effects. *Proceedings of the National Academy of Sciences* 116, 5045-5054 (2019).
4. C. L. Wijnen *et al.*, A complete chromosome substitution mapping panel reveals genome-wide epistasis in Arabidopsis. *bioRxiv*, (2018).
5. V. Calvo-Baltanas *et al.*, Efficient reverse breeding by VIGS-mediated transient crossover reduction. *bioRxiv*, 459016 (2018).
6. M. Ravi, S. W. L. Chan, Haploid plants produced by centromere-mediated genome elimination. *Nature* 464, 615-618 (2010).
7. M. Wędzony *et al.*, in *Advances in Haploid Production in Higher Plants,* A. Touraev, B. P. Forster, S. M. Jain, Eds. (Springer Netherlands, Dordrecht, 2009), pp. 1-33.
8. D. K. Seymour *et al.*, Rapid creation of Arabidopsis doubled haploid lines for quantitative trait locus mapping. *Proceedings of the National Academy of Sciences* 109, 4227-4232 (2012).
9. E. Wijnker *et al.*, Hybrid recreation by reverse breeding in Arabidopsis thaliana. *Nature Protocols* 9, 761-772 (2014).
10. J. A. Birchler, Engineered minichromosomes in plants. *Current Opinion in Plant Biology* 19, 76-80 (2014).
11. M. Ravi *et al.*, A haploid genetics toolbox for Arabidopsis thaliana. *Nature Communications* 5, 5334 (2014).
12. K. Kalinowska *et al.*, State-of-the-art and novel developments of in vivo haploid technologies. *Theoretical and Applied Genetics*, (2018).
13. J. Ren *et al.*, Novel technologies in doubled haploid line development. *Plant Biotechnology Journal* 15, 1361-1370 (2017).
14. T. Ishii, R. Karimi-Ashtiyani, A. Houben, Haploidization via Chromosome Elimination: Means and Mechanisms. *Annual Review of Plant Biology* 67, 421-438 (2016).
15. T. Kelliher *et al.*, Maternal Haploids Are Preferentially Induced by CENH3-tailswap Transgenic Complementation in Maize. *Frontiers in Plant Science* 7, (2016).
16. M. Sanei, R. Pickering, K. Kumke, S. Nasuda, A. Houben, Loss of centromeric histone H3 (CENH3) from centromeres precedes uniparental chromosome elimination in interspecific barley hybrids. *Proceedings of the National Academy of Sciences* 108, E498-E505 (2011).
17. C. M. P. Van Dun, C. L. C. Lelivelt, S. Movahedi. (Rijk zwaan zaadteelt en zaadhandel B.V. (Burgemeester Crezeelaan 40, 2678 KX De Lier, 2678 KX, NL), 2017).
18. R. H. M. O. Den Camp, P. J. Van Dijk, A. Gallard. (Keygene N.V., 2019).
19. A. B. Britt, S. Kuppu, Cenh3: An Emerging Player in Haploid Induction Technology. *Frontiers in Plant Science* 7, (2016).
20. E. Wijnker *et al.*, Reverse breeding in Arabidopsis thaliana generates homozygous parental lines from a heterozygous plant. *Nature Genetics* 44, 467-470 (2012).
21. T. Kelliher *et al.*, One-step genome editing of elite crop germplasm during haploid induction. *Nature Biotechnology* 37, 287-292 (2019).
22. E. S. Lander, D. Botstein, Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* 121, 185-199 (1989).
23. J. H. Nadeau, W. N. Frankel, The roads from phenotypic variation to gene discovery: mutagenesis versus QTLs. *Nature Genetics* 25, 381 (2000).
24. C. Alonso-Blanco, B. Méndez-Vigo, Genetic architecture of naturally occurring quantitative traits in plants: an updated synthesis. *Current Opinion in Plant Biology* 18, 37-43 (2014).
25. W. Bateson, E. Saunders, R. Punnett, C. Hurst, Reports to the Evolution Committee of the Royal Society, Report II. London. *UK: Harrison and Sons*, (1905).
26. H. C. Rowe, B. G. Hansen, B. A. Halkier, D. J. Kliebenstein, Biochemical networks and epistasis shape the arabidopsis thaliana metabolome. *The Plant Cell* 20, 1199-1216 (2008).

27. R. V. L. Joosen *et al.*, Visualizing the genetic landscape of arabidopsis seed performance. *Plant Physiology* 158, 570-589 (2012).
28. B. Méndez-Vigo, J. M. Martínez-Zapater, C. Alonso-Blanco, The Flowering Repressor SVP Underlies a Novel Arabidopsis thaliana QTL Interacting with the Genetic Background. *PLOS Genetics* 9, e1003289 (2013).
29. S. F. Undurraga *et al.*, Background-dependent effects of polyglutamine variation in the Arabidopsis thaliana gene ELF3. *Proceedings of the National Academy of Sciences* 109, 19363-19367 (2012).
30. E. R. Sears, The aneuploids of common wheat. *Research Bulletin 572*, (1954).
31. J. L. Lush, Animal breeding plans. *Animal breeding plans.*, (1943).
32. S. Eberhart, Factors effecting efficiencies of breeding methods. *African soils* 15, 655-680 (1970).
33. T. H. E. Meuwissen, B. J. Hayes, M. E. Goddard, Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics* 157, 1819-1829 (2001).
34. Y. Jiang, J. C. Reif, Modeling Epistasis in Genomic Selection. *Genetics* 201, 759-768 (2015).
35. Y. Zhao, M. F. Mette, J. C. Reif, Genomic selection in hybrid breeding. *Plant Breeding* 134, 1-10 (2015).
36. J. W. Dudley, G. R. Johnson, Epistatic Models Improve Prediction of Performance in Corn. *Crop Science* 49, 763-770 (2009).
37. J. Spindel *et al.*, Genomic Selection and Association Mapping in Rice (Oryza sativa): Effect of Trait Genetic Architecture, Training Population Composition, Marker Number and Statistical Model on Accuracy of Rice Genomic Selection in Elite, Tropical Rice Breeding Lines. *PLOS Genetics* 11, e1004982 (2015).
38. Z. Hu *et al.*, Genomic value prediction for quantitative traits under the epistatic model. *BMC Genetics* 12, 15 (2011).
39. G. De Los Campos, D. Gianola, G. J. M. Rosa, K. A. Weigel, J. Crossa, Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. *Genetics Research* 92, 295-308 (2010).
40. D. Müller, P. Schopp, A. E. Melchinger, Selection on expected maximum haploid breeding values can increase genetic gain in recurrent genomic selection. *G3: Genes|Genomes|Genetics* 8, 1173-1181 (2018).
41. O. Carlborg, C. S. Haley, Epistasis: Too often neglected in complex trait studies? *Nature Review Genetics* 5, 618-625 (2004).
42. S. Salvi, R. Tuberosa, To clone or not to clone plant QTLs: present and future challenges. *Trends in Plant Science* 10, 297-304 (2005).
43. J. Cockram, I. Mackay, in *Plant Genetics and Molecular Biology,* R. K. Varshney, M. K. Pandey, A. Chitikineni, Eds. (Springer International Publishing, Cham, 2018), pp. 109-138.
44. M. C. Ungerer, S. S. Halldorsdottir, J. L. Modliszewski, T. F. C. Mackay, M. D. Purugganan, Quantitative trait loci for inflorescence development in Arabidopsis thaliana. *Genetics* 160, 1133-1151 (2002).
45. R. Mercier, C. Mezard, E. Jenczewski, N. Macaisne, M. Grelon, The molecular biology of meiosis in plants. *Annu Rev Plant Biol* 66, 297-327 (2015).
46. S. Volis, I. Shulgina, M. Zaretsky, O. Koren, Epistasis in natural populations of a predominantly selfing plant. *Heredity* 106, 300 (2010).
47. J. B. Holland, in *Plant Breeding Reviews*. (2001).
48. P. S. Schnable, N. M. Springer, Progress Toward Understanding Heterosis in Crop Plants. *Annual Review of Plant Biology* 64, 71-88 (2013).
49. J. Kuspira, J. Unrau, Genetic analyses of certain characters in common wheat using whole chromosome substitution lines. *Canadian Journal of Plant Science* 37, 300-326 (1957).
50. J. B. Singer *et al.*, Genetic dissection of complex traits with chromosome substitution strains of mice. *Science* 304, 445-448 (2004).
51. L. Wang *et al.*, Hybrid mimics and hybrid vigor in Arabidopsis. *Proceedings of the National Academy of Sciences* 112, E4959-E4967 (2015).
52. E. M. East, Heterosis. *Genetics* 21, 375-397 (1936).
53. Z. J. Chen, Genomic and epigenetic insights into the molecular bases of heterosis. *Nature Reviews Genetics* 14, 471 (2013).
54. A. Fort *et al.*, Disaggregating polyploidy, parental genome dosage and hybridity contributions to heterosis in Arabidopsis thaliana. *New Phytol* 209, 590-599 (2016).
55. P. J. Flood *et al.*, Reciprocal cybrids reveal how organellar genomes affect plant phenotypes. *bioRxiv*,

477687 (2018).
56. D. W.-K. Ng *et al.*, A Role for CHH Methylation in the Parent-of-Origin Effect on Altered Circadian Rhythms and Biomass Heterosis in Arabidopsis Intraspecific Hybrids. *The Plant Cell* 26, 2430-2440 (2014).
57. M. P. Boer *et al.*, A Mixed-Model Quantitative Trait Loci (QTL) Analysis for Multiple-Environment Trial Data Using Environmental Covariables for QTL-by-Environment Interactions, With an Example in Maize. *Genetics* 177, 1801-1813 (2007).
58. F. A. van Eeuwijk, M. C. A. M. Bink, K. Chenu, S. C. Chapman, Detection and use of QTL for complex traits in multiple environments. *Current Opinion in Plant Biology* 13, 193-205 (2010).
59. J. F. Botto, M. P. Coluccio, Seasonal and plant-density dependency for quantitative trait loci affecting flowering time in multiple populations of Arabidopsis thaliana. *Plant, Cell & Environment* 30, 1465-1479 (2007).
60. M. M. Monir, J. Zhu, Dominance and Epistasis Interactions Revealed as Important Variants for Leaf Traits of Maize NAM Population. *Front Plant Sci* 9, 627 (2018).
61. B. E. Huang *et al.*, A multiparent advanced generation inter-cross population for genetic analysis in wheat. *Plant Biotechnology Journal* 10, 826-839 (2012).
62. X. Huang *et al.*, Analysis of natural allelic variation in Arabidopsis using a multiparent recombinant inbred line population. *Proceedings of the National Academy of Sciences* 108, 4488-4493 (2011).
63. P. X. Kover *et al.*, A multiparent advanced generation inter-cross to fine-map quantitative traits in Arabidopsis thaliana. *PLOS Genetics* 5, (2009).
64. J. H. Nadeau, J. B. Singer, A. Matin, E. S. Lander, Analysing complex genetic traits with chromosome substitution strains. *Nature Genetics* 24, 221-225 (2000).
65. A. W. Cowley, R. J. Roman, H. J. Jacob, Application of chromosomal substitution techniques in gene-function discovery. *The Journal of Physiology* 554, 46-55 (2004).
66. R. Koumproglou *et al.*, STAIRS: a new genetic resource for functional genomic studies of Arabidopsis. *Plant J* 31, 355-364 (2002).
67. H. Shao *et al.*, Genetic architecture of complex traits: Large phenotypic effects and pervasive epistasis. *Proceedings of the National Academy of Sciences* 105, 19910-19914 (2008).
68. A. Chen, Y. Liu, S. M. Williams, N. Morris, D. A. Buchner, Widespread epistasis regulates glucose homeostasis and gene expression. *PLOS Genetics* 13, e1007025 (2017).

The Road Not Taken

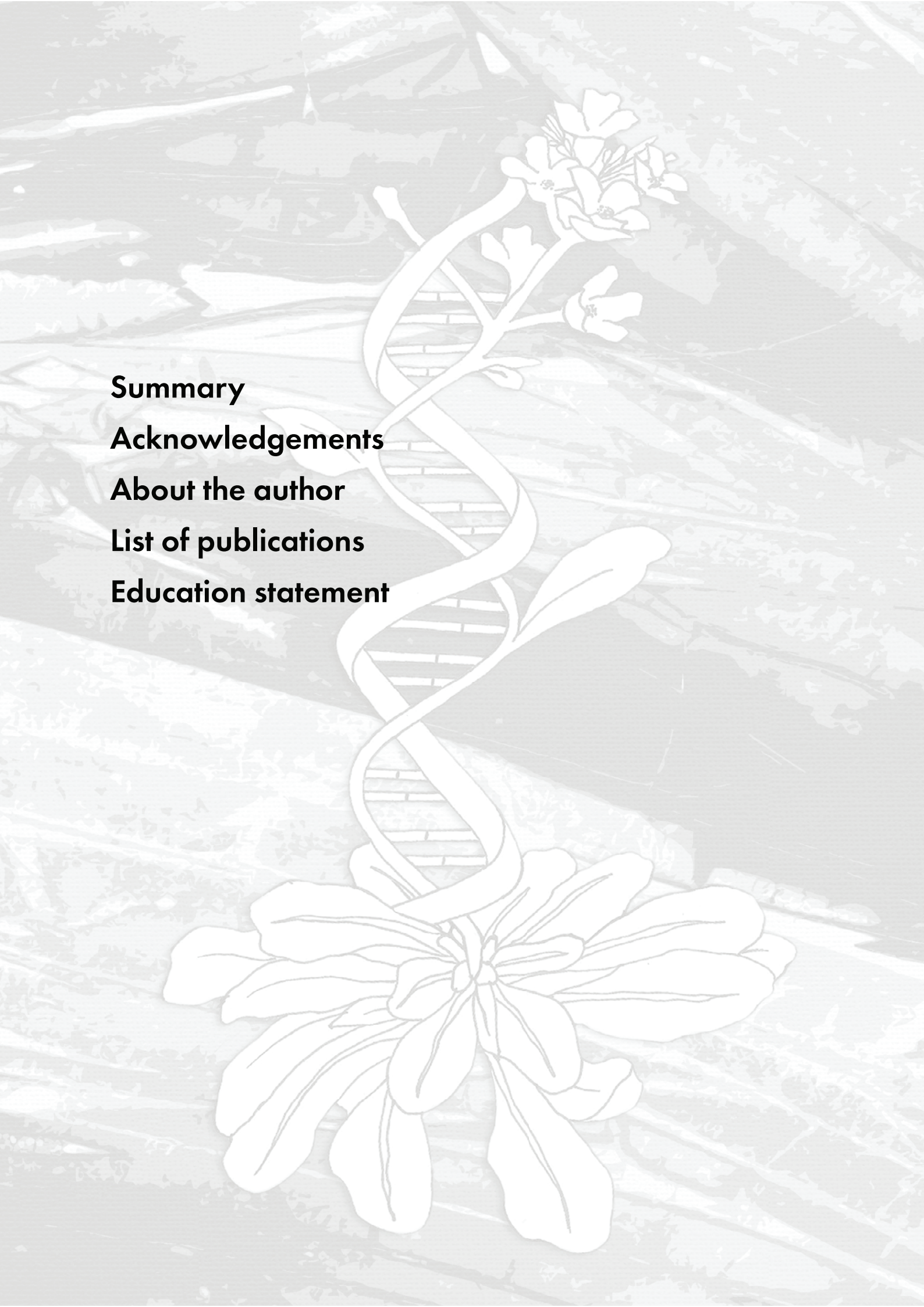
Two roads diverged in a yellow wood,
And sorry I could not travel both
And be one traveler, long I stood
And looked down one as far as I could
To where it bent in the undergrowth;

Then took the other, as just as fair,
And having perhaps the better claim,
Because it was grassy and wanted wear;
Though as for that the passing there
Had worn them really about the same,

And both that morning equally lay
In leaves no step had trodden black.
Oh, I kept the first for another day!
Yet knowing how way leads on to way,
I doubted if I should ever come back.

I shall be telling this with a sigh
Somewhere ages and ages hence:
Two roads diverged in a wood, and I—
I took the one less traveled by,
And that has made all the difference.

Robert Frost



SUMMARY

During meiosis, i.e. the production of the reproductive cells, the number of chromosomes in the cells is halved to ensure that after fertilization the typical quantity of chromosomes is restored. Crossover recombination is an essential step in this process that serves a dual function. On the one hand crossovers ensure the proper segregation of homologous chromosomes and thus ensure the reduction in chromosome number. In addition, crossovers result in the exchange of genetic material between homologous chromosomes. Reshuffling of genetic material is useful for identifying the effects of certain chromosomal segments on phenotypes when specific chromosome segments co-segregate with specific traits. But if this happens at multiple locations and chromosomes simultaneously it can also obscure the detection of those genetic effects. This thesis addresses the question as to how crossover recombination affects the detection of quantitative traits.

The effects of crossover recombination on the detection of quantitative trait loci (QTLs) are explored in this thesis by using doubled haploid (DH) populations in *Arabidopsis thaliana*. As opposed to more traditional mapping populations like recombinant inbred lines, these doubled haploids show lower numbers of crossover events since a doubled haploid experienced only one round of meiosis. In this thesis, some doubled haploids were derived from a modified meiosis because of which they have reduced crossover recombination, or show the complete absence of crossovers. DH derived from a hybrid without crossover recombination are so-called chromosome substitution lines (CSLs). In multiple other species, CSLs are credited to detect QTLs with relative ease due to their simple genetic architecture. Additionally, combinations of CSLs can be used to detect genetic interactions. In this thesis it is explored how genetic analysis can be performed with different sets of CSLs and doubled haploids with reduced crossover recombination. **Chapter 1** further elaborates on the potential use of CSLs and some of the basics of quantitative trait analysis and QTL mapping are explained.

When doubled haploids are made in Arabidopsis, one first produced monoploid (haploid) plants that then give rise to doubled haploids. In **Chapter 2** monoploids and their subsequent doubled haploids were generated, and their phenotypes compared in a genetic analysis to assess the effect of ploidy on their phenotypes. This resulted in detection of QTLs specific for a single ploidy level and QTLs that were expressed in both mono- and diploid generations. The ploidy specific QTLs indicated a different response of the genotypes to the change in ploidy, which was hypothesized to be a response to the sterility of the monoploids. The DH population itself was also used in a QTL mapping approach for flowering time measured in vernalized and non-vernalized conditions. This resulted in the detection of genotype-by-environment

QTLs. Although this confirmed that reduction of crossover recombination (in Doubled haploids) does not *per se* influence genetic mapping, in the following chapters active measures were taken to repress the number of crossover recombinations.

In 2014 reverse breeding was presented as a method in which the parental lines of an F_1 hybrid could be recreated by the combined use of complete suppression of crossover recombination through a stable knock-down of the meiotic recombinase *DMC1* followed by the generation of doubled haploid plants from F1 hybrid. In **Chapter 3** the same approach was used to obtain offspring with no crossover recombination to collect all possible CSL genotypes of a biparental cross. Additionally backcross populations of single chromosome substitution lines (sCSL) with their respective recurrent parent were created. Each of these populations represent a family of lines that segregate for a single chromosome. While the CSLs serve to identify the QTL effects, these single chromosome segregating families were used to finemap major QTL effects. In this thesis it is shown that also for Arabidopsis QTLs can be identified with relative ease in CSLs. Additionally, by acquiring all the different combinations of CSL genotypes, one can also look for genetic interactions between chromosomes. Genetic interaction effects are assumed to occur when the phenotype of an individual in which two independent loci are substituted cannot be predicted based on the individual effects of those two loci. In **Chapter 3** such interaction effects between chromosomes are identified for two traits (flowering time and main stem length). Such genetic interaction effects are shown to have effect sizes equal to typical QTLs effect sizes. This demonstrates the advantage of discovering possible genetic interactions using such CSL populations.

The experiment to detect epistasis (i.e. genetic interactions) using a panel of all possible CSL genotypes in **Chapter 3** was limited to only two phenotypes. This did not allow generalizations on how frequent epistasis is observed. This question was in part answered in **Chapter 4**, where a shotgun-proteomics approach was used to detect main effects and interaction effects between chromosomes for the abundance of proteins. Especially a small but genetically diverse panel of CSLs allows the detection of genetic effects at the proteome level, which was so far not technologically feasible in a cost-efficient approach. Only a subset of CSLs was used, which illustrates that a limited set of fourteen genotypes with single and double substitutions in a recurrent background can be used to detect genetic effects. Using protein abundancy as a phenotype, more than a thousand QTLs divided over the different chromosomes were identified. Especially chromosomes 2 and 5 were identified as contributors to the proteome variation. Furthermore, it was shown that the abundance of approximately 20% of the measured proteins was significantly dependent on the presence of a combination of two chromosomes (i.e. an interaction effect). Furthermore, for several proteins their abundance was not influenced by the

allelic state of any single chromosome, but they were only significantly influenced by a combination of chromosomes. This shows that interaction effects can also be detected when a limited number of CSLs are used.

The previous chapters deal with confirming the utility of CSLs, **Chapter 5** addresses several practical difficulties of the application of reduced crossover recombination to advance plant breeding practice. Because proper homolog segregation is compromised in the absence of crossover recombination, plants produce high numbers of non-viable spores and have a very low fertility. In **Chapter 5** the effect is studied of the knock down of *MSH5* instead of *DMC1*. With a dysfunctional MSH5 crossover recombination is not completely absent but suppressed to retain still a few crossovers per meiosis. Therefore, proper homolog segregation occurs more often, and plants remain more fertile. Additionally, instead of stable suppression via transformation of a parental line, virus-induced-gene-silencing (VIGS) is used to transiently downregulate the gene MSH5 in the meiosis of a hybrid directly. Both revisions together ensure transgene-free progeny in only three generations with a few or no recombination events per offspring. Using such progeny, the original starting hybrid could be recreated using complementing non-crossover offspring which did not differ phenotypically from the original hybrid. Additional intercrosses were made between non-crossover and offspring containing few crossovers that were complementary for most of the genome to produce hybrids with small homozygous segments. In general, such near-full hybrids did not differ phenotypically from the original hybrids with only few exceptions. This suggests that one could improve hybrids directly by fixating specific segments containing for instance recessive QTLs. The improvements to reverse breeding make this breeding approach more efficient and flexible for application to higher chromosome number species.

In **Chapter 6** a discussion follows on the previous chapters. Also new and additional experiments are proposed based on the observations described in the different chapters. It is argued that a complete CSL panel with all possible CSL genotypes can give a better overview of the genetic architecture controlling quantitative traits, especially when multiple interacting QTLs are to be expected. Smaller panels of CSLs might be informative enough for the purpose of genetic mapping in high chromosome number species, and practical considerations should include the type of genetic effects one wants to identify and the known genetics of the considered trait. Taken together, the work in this thesis shows that with different approaches reduced crossover recombination can additional advantages on the analysis of quantitative traits.

ACKNOWLEDGEMENTS

My first gratitude is addressed to my entire promotion team: **Joost**, **Fred**, **Martin** & **Erik**. Together you have given me the opportunity to learn about quantitative genetics, plant breeding, statistics and science in general. There have been many valuable lessons during the past years that I would not have been able to learn without your support. **Joost**, you gave me guidance during the PhD. We have often discussed the possible directions we should take for the different chapters and which subjects we should explore first. However, you always considered this also as my project and therefore gave me liberty to explore my own ideas. I will always remember your openness and honesty, there were no secrets or topics we could not discuss during our meetings or the barbeques at your place with your family and team. **Fred**, I enjoyed the moments you managed to take the time to explain me in detail how a specific statistical analysis should be performed. I appreciate that you ensured that a PhD program was more than working office hours. You made sure I was engaged with other PhD students from your department through organising a symposium, a PhD-day, and going out for a drink from time to time. **Martin**, I still remember the first time I came into your office; you just came back from New York after finishing the famous New York Marathon. I had only recently started my new running hobby and was not yet going for distances above fifteen kilometres. Since that first meeting we shared our passion for running almost every time we had a work discussion. I felt those were pleasant additions to the technical discussion on statistics and data analysis. You were always available for providing quick feedback or for answering my one-short-questions that usually had multiple long answers. **Erik**, initially you were not part of my supervision team. At the moment of writing, it is unbelievable to foresee how this thesis would have come to be without you being a part of my supervision team. This thesis builds on your previous work but you were only sparsely in Wageningen until half way through my PhD. Since your return to the Laboratory of Genetics you have been involved with basically every aspect and we worked closely together on most of the chapters in this thesis. Sometimes I got overwhelmed by your creative mind which is not stopped by the current limitations in science. Discussions could go in every direction and it sometimes left me more confused than when we started, but it also helped to improved my critical way of thinking. To explore a world with a scientific approach, we do not always need to limit ourselves by our present knowledge. **To all four I would like to say thank-you** for guidance and providing your assistance during this thesis. Thanks for your trust in me. Although I took some of the liberty you provided to make this work my own, I still hope the result resembles the initial ideas you had in mind.

I would also like to express my gratitude to **RijkZwaan** and especially **Rob**, **Cilia**, **Evert** & **Kees**, who were willing to invest in this project and share their feedback during the many meetings we had over the years. **Bastiaan**, without your super-fast work in the RijkZwaan lab I am certain that this thesis would not have been possible in the way it is.

I have spent quite some time in the Laboratory of Genetics and as the cliché goes, I have seen a lot of people come and go, not just within Genetics but also in other departments. I would like to thank all of you (hopefully without forgetting someone): **Aina, Alex, Andy, Anna, Anneloes, Arjan, Bart, Bas, Ben, Bernard, Bwalya, Carla, Claudio, Corrie, Daniela, Duur, Eric, Ernest, Eveline, Florien, Fons, Francesco, Frank, Gabriëlla, Gil, Hans, Jan, Jelle, Jianhua, Jitpanu, Job, Johanna, Joost van de Heuvel, Jordy, José, Juriaan, Justin, Kim, Klaas, Krithi, Laurens, Lennart, Luuk, Lydia, Maarten, Maggie, Marcela, Marcin, Margo, Margreet, Marieke, Marijke, Mariska, Marjon, Mark Aarts, Mark Zwart, Matthijs, Mina, Padraic, Paola, Peter, Petra, Philip, Phuong, Pingping, Ramon, Raphael, René, Rik, Robert, Roel, Ross, Roxanne, Sabine, Sarah, Sijmen, Tania, Tina, Tom, Twan, Valeria, Vanesa, Vincent, Wytske, Xianwen, Yanli.** The lab of Genetics has been a very nice and pleasant department and I would like to thank all the people that have made this PhD journey more fun during the nice times and more bearable during tough times. I am happy to have enjoyed the many nice moments such as the quizzes, dinners and 'uitjes' organised by the TGAC or the discussions during lunchbreaks, but we also shared some tough moments. Basically, there are too many stories that spring to mind when looking over this list of people and I prefer to talk about those over a beer instead. Still, there are a number of people I would like to thank here in a more personal way:

Bas, I would like to express my appreciation to you as the department chairholder. You are always very supportive of the rest of the department, so that they are able to perform their and your passion for genetics and evolution. You are managing the position as chairholder very well and I expect that under your guidance the lab will continue to perform at a high scientific level. Thank-you for your support, even when I was overextending my stay within the department. Also **Wytske**, you have been a great support, and probably I do not even know half of the things you arrange for the department to run smoothly. **Hans & Maarten**, both of you retired from your position as professor at the Laboratory of Genetics during my PhD, but you also show that retirement does not mean less passion for science. Curiosity and enthusiasm still trigger your bright minds and I hope to continue my life with similar virtues.

Although officially I was part of Biometris and the Laboratory of Genetics, I spend most of my time neither in the Biometris department nor in the lab of the Genetics

department: Most of my work was performed in the office and Unifarm facilities of Genetics. In the greenhouses and climate chambers my plants were handled by a number of dedicated people from the Unifarm facilities that deserve to be mentioned here because without proper care of my plants, none of the experiments could have been performed and no data could have been generated for further analysis. Among those who I have bothered most with ignorant questions on where to find stuff or how to grow plants are **Gerrit**, **Taede**, **Rohan & Bert**. Thank-you for taking care of those little plants and for always being ready to help out!

José, **Laurens**, you were both part of the early CSL team. I appreciate all the work that was done before I started and thank you for answering all the questions I had on the protocols and on the history of all the different seed bags and excel files that you created. **José**, you were my guide through the first weeks in Genetics and the teacher for creating the CSLs. You taught me how to grow and cross the plants and I still remember the pains I had in my hands from the first crossings. **Laurens**, although you already left the Laboratory of Genetics when I started you were only one flight of stairs away, so you met us often enough to stay updated on the proceedings of the CSLs. **Frank**, you have been a great help and support during the PhD. There have been many times I was performing experiments that could not have been performed without your assistance, e.g. creating the setup of the climate chambers, sowing, phenotyping or harvesting. Even for lab work, for which I was originally trained, you have supported me (and my students) because my lab skills had gone rusty due to time spent on data analysis and writing. I also enjoyed a lot the times we were searching for Arabidopsis together with the students of the MEE-course. It was a lot of fun! Your life changed quite a bit during the years of my PhD, and I hope to visit you again soon at your family farm. Thanks for everything.

Corrie, thanks for all the holiday advice. Although you always tell me we travel at a different pace, I think we both enjoy the mountains and nice trails in the forests. **Mark**, thank-you for all advices during lunchbreaks, I am very much looking forward to all the innovation the new phenotyping facilities will provide. **Fons**, thanks for all the trust in me and allowing me to set up my own experiment in the GATC course. Even though my excel tables were sometimes a bit confusing, I hope the CSLs continue to be a nice tool for explaining quantitative genetics to the students. Also during the MEE-course I was always surprised with all the different topics that the students could choose from. **Bart**, it is inspiring to see how well you manage all the students, while still allowing them to do the experiments they would like to do. **Sijmen**, thanks for your encouragements, especially in the last months of my PhD.

The group of Joost was never very big and it was split between Plant Physiology and Genetics but this provided nice additional input via the weekly meetings. **Rik**,

although you were more or less finishing your PhD when I was just starting, we met a couple of times due to your PostDoc positions and photography hobby. Thanks for showing that it is possible to have a time-consuming hobby next to the PhD. **Emilie**, you were always willing to help out whether it was listening to a complaint I had or providing some additional climate room space. You were willing to be critical at times when no one else was. Thanks, I have appreciated a lot! **Juriaan**, you started your own PhD adventure, I have no doubts that you will succeed, your perseverance and integrity are the cornerstones of a good scientific basis. **Yufeng, Mengfan, Joram, Claudio & Ruth,** all of you helped me with my research while performing your MSc thesis at the Laboratory of Genetics. Although unfortunately almost none of the work that has been done by you ended up in my thesis, I still believe there was a lot of benefit to those projects. You all taught me again how valuable the process of teaching and supervising is and hopefully the projects will be continued at some point by new students. It makes me also proud to know that at least half of you is currently pursuing his/her own PhD, apparently I did not scare you.

Valeria, **Vanesa**, **Ramon**, we shared most of our PhD life together. I cannot conclude otherwise than that it has been an intense experience and it would not have been the same without you. **Valeria**, you are an excellent scientist and a great teacher, you have helped me to understand myself a little bit better. **Vanesa**, you can be contagiously passionate about many things in life, be it your work, your environmental activism or whatever funny thing you find interesting. **Ramon**, our PhD projects have been intertwined from the start. We have shared difficult moments and pleasant surprises in science and in our personal lives. A PhD can bring a lot of stress and without you, I am certain I would not have been able to cope with all of it the way I did. I have appreciated your presence, and I envy your positive go-getter attitude. Every week you had fresh thoughts, or you had new ideas to organise some strange Droef event. I am sure that when you restart your PhD activities, you will be able to finish it, and when you do, I will try to help however I can. Many thanks to all three of you for being there during my PhD. Although we are all having very different lives now I hope we can continue to stay in contact.

Lennart & Aranka, first, thanks again for accepting to be my paranymphs. It is a great honour to have you at my side during the defence. **Lennart**, I think it took a while before we both realized that we were both doing quite some intensive endurance sport activities next to the PhD. I was running (ultra)marathons in the mountains and you were participating 24-hour skate events. It is truly amazing to see what you have been able to achieve in the past couple of years. For me it was always these nice conversations that coloured my days, they became pleasant distractions. Thanks for your advice when I was on mushroom hunts (also **Alex**, **Anna**, **Frank** & **Maurits**). Lennart, you are also a very clever man and I am sure you will also finish

with an interesting PhD. **Aranka**, you have been the best friend for me during this adventure. I met you during the MSc, and since then we have shared many coffee- and lunchbreaks together. You know I have had a lot of ups and downs during this process but I could share all of my concerns with you. Thanks for listening and for all the advises during coffee breaks. You will always be a welcome friend to visit wherever I am.

Twan & **Jan**, it has been 10 years since I first went to Wageningen and started my BSc internship with you. You were then and still are very pleasant people to work with. I am glad I took the step 10 years ago to knock on your door, and I am happy Jan came to me to collaborate for a ZonMW project. **Twan**, I really liked to be back in the proteomics lab, which brought back many nice memories, but I hope our current work will lead to something more than the current thesis chapter. **Jan**, whenever you want to go for a nice bike ride in Limburg, you are more than welcome to visit.

Also I want to thank some of my other friends that have contributed to this PhD by providing support and trust. **Pedro, Nely, Gus, Pame, Chris, Wilco, Grace,** I can honestly say that you guys have enriched my life similar to how the classical Dutch cuisine (Pame, is there anything like Dutch cuisine?) has been enriched with foreign spices. It feels like you bring bright colours and tastes into my life whenever I meet you, and I enjoy all the moments we manage to spend together.

Bart, **Paul** & **Erik**, bedankt voor al jullie steun gedurende de afgelopen jaren. Jullie hebben vanaf het begin vertrouwen erin gehad dat ik een PhD succesvol zou afronden. Ik vind het geweldig dat we nog steeds ieder jaar proberen samen wat leuks te gaan doen. Het heeft wat langer geduurd dan verwacht maar dit jaar organiseer ik het feestje.

Zoals al een paar keer is genoemd, ik heb tijdens mijn PhD ook aardig wat kilometers gemaakt, zowel op de fiets, als hardlopend. Het fietsen gebeurde vooral in en rondom Zuid-Limburg samen met de '*Wielder Klimgeiten*'. **Jef & Marcel** bedankt voor de onvergetelijke fiets reizen naar Noorwegen (Viking Tour) en Frankrijk (Marmotte). Ook de wedstrijd hardloopgroep van Pallas '67, of '*Track & Trail Busters*' onder leiding van **Jeroen** heeft een grote bijdrage gehad aan de totstandkoming van dit proefschrift. Dankzij je trainingsschema's kon ik mijn PhD werk beter structureren. Daarnaast boden jullie zeker in het laatste jaar steeds vaker een luisterend oor tijdens de lange duurlopen. Hardlopen is een hele gezonde manier om de gedachte te verzetten, maar samen met jullie werd het ook heel prettig. Bedankt allen, dat jullie de passie voor duursporten met mij hebben gedeeld in deze periode.

Maurice, Jolanda, Meike, Peter, Marthe, Tess, Toby, Stefan, Emmy, Sjoerd, Bart, Rachelle, Zoë, Robin, Noa, Mark, Ryanne, toen ik in 2013 begon met de

PhD voetbalde ik nog bij V.V. Wijlre, had ik slechts één keer deelgenomen aan een 10 kilometer hardloopwedstrijd en waren er nog geen kinderen geboren in onze vriendengroep. Het geeft aan dat er in deze periode veel is veranderd, en ondanks de afstand die ontstond door de verhuizing zijn we altijd leuke dagen en weekenden samen blijven beleven. Ik ben blij dat ik zo een groep vrienden heb waarmee we nog steeds de mooie maar ook de minder leuke momenten in het leven samen kunnen delen.

Dan uiteindelijk mijn **familie**, ik bedank jullie allemaal voor de interesse in mijn werk. **Wino** & **Melania**, you both went before me along this path of trial and error and with great respect I see your passion for science. It is encouraging to see such examples with whom I could talk for hours about my PhD. Unfortunately we can not spend more time together, but Switzerland will always feel a bit closer as long as you are there, thanks! **Mike** & **Lisa**, ik heb jullie niet eerder genoemd, maar dat had makkelijk gekund tussen de vrienden en mede duursporters, jullie hebben gezorgd dat ik van tijd tot tijd ook mijn gedachten kon verzetten. Samen sporten, naar de stad, een avondje uit eten, of thuis een spelletje spelen, bedankt voor al het plezier dat zorgde voor de ontspanning. **Pap** & **mam**, bedankt voor alle liefde en steun die jullie mij in mijn leven hebben gegeven. Ik heb altijd een prachtige plek om terug naar huis te gaan. Bedankt dat ik van jullie mag zijn wie ik ben.

Lieve **Wendy**, jou liefde stelt mij in staat om te doen wat ik leuk vind, alles wat ik heb bereikt in de afgelopen jaren was mogelijk door jou. Je bent getuige geweest van alle moeite die het mij heeft gekost, alle twijfels die ik had, en iedere traan en lach die ik er voor over had. We hebben ondertussen al behoorlijk wat avonturen samen meegemaakt en we hebben ook deze periode met z'n tweeën beleefd, maar jij hebt voor mij gezorgd, en ik wil graag voor jou zorgen. Ik kijk uit naar een mooie toekomst voor ons samen, waarin we kunnen zeggen dat ons proefschrift af is.

CURRICULUM VITAE

On April 24th 1987 Cristian Lucas Wijnen was born in Wijlre in the middle of the Dutch '*Heuvelland*' in the South of The Netherlands. His time at primary school "Op de Tien Bunder" was accompanied by enjoying the country life and the welcoming home of his grandparents. After primary school, he went to the secondary school at the Sophianum in Gulpen. Here he followed the HAVO Nature & Health program and graduated in 2004.

Both sports and biology had his interest so he decided to subscribe for studying Physical Therapy at Hogeschool Zuyd in Heerlen. After one year he decided that physiology and molecular cell biology were the most interesting classes, so he chose to switch to studying Applied Sciences with a specialisation in Molecular Life Sciences and a minor in Educational Skills at the Fontys Hogescholen in Eindhoven. His first internship at Maastricht University was in the Clinical Genomics lab while his second internship in 2009 was in the proteomics lab of the Bioscience department of Plant Research International in Wageningen. It was here that he (re)found his passion for biology and plants.

Directly after finishing his BSc degree in August 2010, he continued to work at the Applied Sciences program as a research and educational assistant. While working in Eindhoven he also was a substitute teacher at the Regional Education Centre for a short period. In March 2011 he decided to subscribe for the master studies Plant Biotechnology at the Wageningen University & Research. Shortly afterwards he came in contact with Nunhems Netherlands B.V. and decided to take the opportunity of getting a couple of months of experience in a vegetable breeding company.

In September 2011 he started his studies of Plant Biotechnology with a specialisation in Molecular Plant Breeding and Pathology and a minor in Plant – Insect Interactions. During this two-year study, he performed a thesis at the Laboratory of Entomology where he performed a genome wide association study to dissect the genetic basis of Arabidopsis – specialist insect interactions under the supervision of Dr. Davila Olivas and prof. Dr. van Loon. His second thesis he performed in Cologne, at the Max Planck Institute for Plant breeding research. This thesis encompassed a quantitative trait locus analysis of flowering time and photoperiod sensitivity in tomato at the Adaptive Genetics & Genomics department and was supervised by Dr. Müller and Dr. Jiménez-Gómez.

November 2013 marked the beginning of his work on his PhD thesis at the Laboratory of Genetics and Biometris departments of Wageningen University & Research. The result of his PhD are described in this thesis. Currently he is working again for Nunhems Netherlands B.V. as an associate scientist pepper at the Molecular Assisted Breeding department.

LIST OF PUBLICATIONS

Wijnen, C. L. and J. J. B. Keurentjes (2014). "Genetic resources for quantitative trait analysis: novelty and efficiency in design from an Arabidopsis perspective." Current Opinion in Plant Biology 18: 103-109.

Müller, N. A., **C. L. Wijnen**, A. Srinivasan, M. Ryngajllo, I. Ofner, T. Lin, A. Ranjan, D. West, J. N. Maloof, N. R. Sinha, S. Huang, D. Zamir and J. M. Jiménez-Gómez (2015). "Domestication selected for deceleration of the circadian clock in cultivated tomato." Nature Genetics 48 (1): 89-93.

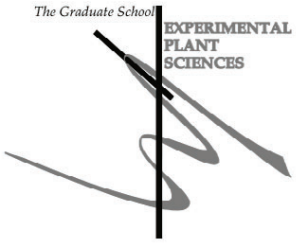
Fort, A., P. Ryder, P. C. McKeown, **C. Wijnen**, M. G. Aarts, R. Sulpice and C. C. Spillane (2016). "Disaggregating polyploidy, parental genome dosage and hybridity contributions to heterosis in Arabidopsis thaliana." New Phytologist 209 (2): 590-599.

Davila Olivas, N. H., W. Kruijer, G. Gort, **C. L. Wijnen**, J. J. A. van Loon and M. Dicke (2017). "Genome-wide association analysis reveals distinct genetic architectures for single and combined stress responses in Arabidopsis thaliana." New Phytologist 213 (2): 838-851.

Wijnen, C. L., R. Botet, J. van de Belt, L. Deurhof, H. de Jong, B. C. B. de Snoo, R. Dirks, M. P. Boer, F. A. van Eeuwijk, E. Wijnker and J. J. B. Keurentjes (2018). "A complete chromosome substitution mapping panel reveals genome-wide epistasis in Arabidopsis." biorxiv.org/content/10.1101/436154v1.

Calvo-Baltanas, V., **C. L. Wijnen**, N. Lukhovitskaya, C. B. de Snoo, L. Hohenwarter, H. de jong, A. Schnittger and E. Wijnker (2018). "Efficient reverse breeding by VIGS-mediated transient crossover reduction." biorxiv.org/content/10.1101/459016v1.

EDUCATION STATEMENT OF THE GRADUATE SCHOOL EXPERIMENTAL PLANT SCIENCES



Issued to: Cristian Lucas Wijnen
Date: 14 October 2019
Group: Laboratory of Genetics & Biometris
University: Wageningen University & Research

1) Start-Up Phase	*date*	*cp*
► **First presentation of your project**		
Powerful new resources for studying the genetic basis of complex traits in Arabidopsis	21 Jan 2014	1.5
► **Writing or rewriting a project proposal**		
► **Writing a review or book chapter**		
Wijnen, C.L. & Keurentjes, J.J.B. (2014) Genetic resources for quantitative trait analysis: novelty and efficiency in design from an Arabidopsis perspective. *Current Opinion in Plant Biology* https://doi.org/10.1016/j.pbi.2014.02.011	20 Mar 2014	3.0
► **MSc courses**		
Subtotal Start-Up Phase		4.5

2) Scientific Exposure	*date*	*cp*
► **EPS PhD student days**		
EPS PhD student days "Get2Gether", Soest, NL	29-30 Jan 2015	0.6
EPS PhD student days "Get2Gether", Soest, NL	28-29 Jan 2016	0.6
EPS PhD student days "Get2Gether", Soest, NL	9-10 Feb 2017	0.6
EPS PhD student days "Get2Gether", Soest, NL	15-16 Feb 2018	0.6
► **EPS theme symposia**		
EPS Theme 4 symposium 'Genome Biology', Wageningen, NL	13 Dec 2013	0.3
EPS Theme 1 symposium 'Developmental Biology of Plants', Wageningen, NL	24 Jan 2014	0.3
EPS Theme 4 symposium 'Genome Biology', Wageningen, NL	3 Dec 2014	0.3
EPS Theme 4 symposium 'Genome Biology', Amsterdam, NL	15 Dec 2015	0.3
EPS Theme 4 symposium 'Genome Biology', Wageningen, NL	16 Dec 2016	0.3
► **Lunteren Days and other national platforms**		
Annual meeting 'Experimental Plant Sciences', Lunteren, NL	14-15 Apr 2014	0.6
Annual meeting 'Experimental Plant Sciences', Lunteren, NL	13-14 Apr 2015	0.6
Annual meeting 'Experimental Plant Sciences', Lunteren, NL	11-12 Apr 2016	0.6
Annual meeting 'Experimental Plant Sciences', Lunteren, NL	10-11 Apr 2017	0.6
Annual meeting 'Experimental Plant Sciences', Lunteren, NL	9-10 Apr 2018	0.6

► **Seminars (series), workshops and symposia**		
Seminar: Prof. Dr. Eric Schranz: Whole genome duplications as drivers of evolutionary innovations and radiations?	21 Nov 2013	0.1
Symposium: All-Inclusive Breeding: Integrating hightroughput sciences	16 Oct 2014	0.3
Seminar: Prof. Dr. George Coupland: Seasonal flowering in annual and perenial plants	19 Jan 2015	0.1
Symposium: Future for cytogenetics in plant genomics and breeding	1 Oct 2015	0.2
Seminar: Prof. Dr. Alain Tissier: Insight into the inner workings of a metabolic cell factory	18 Mar 2016	0.1
Symposium: Zoology: Genotype-phenotype map: From model systems to ecosystems	23-24 Nov 2017	0.6
Symposium: Genotype to Phenotype Modelling of Plant Adaptation	16 Nov 2017	0.3
► **Seminar plus**		
► **International symposia and congresses**		
EUCARPIA Section Biometrics in Plant Breeding, Wageningen, NL	9-11 Sep 2015	0.9
26th International Conference on Arabidopsis Research, Paris, France	5-9 Jul 2015	1.4
► **Presentations**		
Annual meeting 'Experimental Plant Sciences', Lunteren, NL (Talk & Poster)	14 Apr 2015	2.0
26th International Conference on Arabidopsis Research, Paris, France (Talk)	8 Jul 2015	1.0
Annual meeting 'Experimental Plant Sciences', Lunteren, NL (Talk & Poster)	11 Apr 2017	2.0
Symposium Genotype to Phenotype Modelling of Plant Adaptation (Talk)	16 Nov 2017	1.0
► **IAB interview**		
► **Excursions**		
Subtotal Scientific Exposure		16.9

3) In-Depth Studies	date	cp
► **Advanced scientific courses & workshops**		
GenStat QTL course, Wageningen, NL	2-4 Sep 2013	1.0
Introduction to R for Statistical Analysis, Wageningen, NL	19-20 May 2014	0.6
EMBL Next Generation Sequencing: Whole Genome Sequencing Library Preparation, Heidelberg, Germany	20-21 Apr 2015	0.6
Genome Assembly, Wageningen, NL	28-29 Apr 2015	0.6
Data Analyses and Visualizations in R, Wageningen, NL	11-12 May 2017	0.6

► **Journal club**		
Botanical Genetics literature discussion group	2013-2017	2.0
Multi-parent population discussion group (Genetics & Biometris)	2015	1.0
► **Individual research training**		
Subtotal In-Depth Studies		6.4

4) Personal Development	*date*	**cp**
► **General skill training courses**		
Competence Assessment, Wageningen, NL	25 Mar 2014	0.3
Data Management Planning, Wageningen, NL	15 Feb 2016	0.4
WGS Course Career Orientation, Wageningen, NL	7-28 Mar 2017	1.5
WGS Course Communication with the Media and the General Public, Wageningen, NL	9 Nov - 11 Dec 2017	1.0
Scientific Artwork, Wageningen, NL	19-20 Mar 2018	0.6
► **Organisation of meetings, PhD courses or outreach activities**		
Symposium All-Inclusive Breeding: Integrating highthroughput sciences, Wageningen, NL	16 Oct 2014	1.5
► **Membership of EPS PhD Council**		
Subtotal Personal Development		5.3

TOTAL NUMBER OF CREDIT POINTS*	**33.1**

Herewith the Graduate School declares that the PhD candidate has complied with the educational requirements set by the Educational Committee of EPS with a minimum total of 30 ECTS credits.

** A credit represents a normative study load of 28 hours of study.*

The research described in this thesis was conducted at the Laboratory of Genetics, Wageningen University & Research, the Netherlands. This work has been financially supported by the Netherlands Organisation for Scientific Research under grant number STW-12425, for which additional support was received from Rijk Zwaan B.V..

Cover artwork by Wendy Simon
Layout design by Iliana Boshoven Gkini | AgileColor.com
Printed by GVO drukkers & vormgevers B.V. on FSC-certified paper