Phenotypic variation in plants

Roles for epigenetics

Lauss, K.

Publication date
2017
Document Version
Other version
License
Other

Chapter 2

QTL^{epi} mapping in *Arabidopsis thaliana*

Kathrin Lauss[1] and Joost J.B. Keurentjes[2]

[1] University of Amsterdam, Swammerdam Institute for Life Sciences, Science Park 904 1098XH Amsterdam, The Netherlands.

[2] University of Wageningen, Laboratory of Genetics, Droevendaalsesteeg 1, 6708PB Wageningen, The Netherlands.

Abstract

While DNA sequence variation is known to be a major driver of phenotypic divergence, epigenetic variation has long been disregarded. One reason for that was the lack of suitable tools. The creation of epigenetically divergent but otherwise largely isogenic Arabidopsis populations has now alleviated some of these constraints. Epigenetic recombinant inbred line (epiRIL) populations allow examining the effects of epigenetic variation on phenotypes. In addition, epiRILs enabled the development of epigenetic quantitative trait loci (QTL^{epi}) mapping, an approach to identify causal epigenetic factors. Here we describe the subsequent steps of QTL^{epi} mapping in a broad sense, from the creation of epigenetically divergent populations through plant phenotyping to the identification of causal genes underlying particular phenotypes in Arabidopsis.

1 Introduction

Intraspecific natural variation (from here on termed natural variation) is defined as the wide phenotypic and adaptive diversity occurring in nature within a single plant species and caused by genetic differences.

Natural variation has been effectively exploited to associate genotypic divergence with phenotypic trait variation and has been recognized as a valuable source for agricultural crop improvement [3]. Phenotypic traits can be categorized qualitatively and quantitatively. Qualitative trait variation can be classified in distinctly defined phenotypic categories, is often monogenic regulated and the influence of the environment on the trait is subtle (e.g. resistance to particular pathogens) [3]. Quantitative traits, on the other hand, are polygenic, stronger influenced by the environment and typically show a wide continuous variation in

phenotypes (e.g. the onset of flowering) [3,4]. Natural variation occurs for both qualitative and quantitative traits. Studying a qualitative trait is rather straightforward and a phenotype can be easily associated with the causal gene by using forward or reverse genetic approaches. By contrast, a quantitative trait will display a continuous distribution of phenotypic values for different genotypes, thus, more sophisticated tools are needed to link quantitative traits to their causal genes [3].

One way to identify causal loci underlying quantitative trait variation is by using segregating populations for genetic mapping to detect quantitative trait loci (QTLs). These approaches make use of sequence polymorphisms between lines that are used as genetic markers to label the parental origin of genomic fragments in progeny of crosses between those lines. Screening a large population of lines with a varying composition of "marked" parental genomic fragments basically allows monitoring co-occurrence, or co-segregation, of particular fragments with a specific phenotypic trait value.

Widely used populations for QTL mapping are recombinant inbred lines (RILs). RILs are commonly generated by crossing genotypically (and phenotypically) divergent inbred lines and propagation by single seed descent until homozygosity to obtain a set of immortal lines harboring different genomic introgressions from their parents. However, classical RIL populations do not take into account epigenetic variation.

Variation in epigenetic marks includes DNA methylation (at cytosine-residues) and histone modifications, which are known to regulate gene expression in a wide range of eukaryotic organisms, including plants [7,101,102]. It has been suggested to distinguish between transgenerational epigenetic memory of gene expression states, which is inherited across generations, and transient epigenetic changes induced by developmental or environmental stimuli [103,104]. Examples for transgenerational epigenetic memory in plants are epialleles: allele

variants that consist of an identical DNA sequence but which are differentially transcribed due to differences in their epigenetic modifications [34,35,105]. Epialleles being causative for particular phenotypes have been described in various plant species [11–15], *e.g.*, the *Flowering Wageningen* (*FWA*) locus in Arabidopsis occurs in two epiallelic forms: a repressed state associated with extensive DNA methylation and a demethylated, highly transcribed state that causes a late flowering phenotype [36]. Hence, natural variation is, in addition to genetic factors, also regulated by epigenetic factors. That said, non-Mendelian segregation patterns of specific epigenetically regulated phenotypes can also be observed [81]. For instance, epialleles can undergo so called paramutation or trans-chromosomal methylation and demethylation events during hybridization [79,81]. Thereby, one epiallele acquires the epigenetic profile (often DNA hypermethylation) of the other epiallele. Epialleles and paramutation phenomena thus exemplify the involvement of epigenetic variation in shaping plant phenotypes.

Currently, resources for studying and quantifying epigenetic variation and its association with phenotypic variation are rapidly emerging. Epigenetic recombinant inbred lines (epiRILs) are among the most informative of those tools. Analogous to RILs being used to identify causal genomic loci explaining trait variation, epiRILs can be used to detect causal loci in the epigenome (QTL^{epi} mapping).

Here, we explain the use of epiRILs for the identification of epigenetic patterns contributing to quantitative trait variation. We address aspects of experimental design, plant phenotyping and data analysis.

2 Materials

2.1 Epigenetic recombinant inbred lines (epiRILs)

In order to study epigenetic variation and its influence on traits independently from genetic variation, epigenetic recombinant inbred lines (epiRILs) are a powerful experimental system. An epiRIL population is conceptually the same as a generic recombinant inbred line (RIL) population. The main difference is that RILs acquired a mosaic of different DNA sequence introgressions from their parents while epiRILs acquired a mosaic of epigenetic patterns from their isogenic parents.

A RIL population is generated by creating a hybrid from two genetically (and phenotypically) distinct inbred lines of interest followed by single seed descent from the F2 generation onwards until at least the F8 generation. This process results in an accumulation of several recombination events across the chromosomes and through the many generations of inbreeding those lines reach (near) full homozygosity.

The creation of epiRILs is very similar except that the inbred lines used as parents are isogenic apart from one parent carrying a mutation with a strong effect on epigenetic profiles. Rounds of backcrossing and selfing until the F8 result in a population of genetically nearly identical lines, which are segregating for epigenetic variation. Nonetheless, remaining sources of genetic variation in epiRILs could be transposable elements which lost their repressive epigenetic marks and became active again [20]. Therefore, in epiRILs it is useful to distinguish (phenotypic or molecular) effects induced by transposon insertions from those induced by epigenetic variation.

Confirmed strategies to create epiRIL populations

To date, two epiRIL populations have been generated in *Arabidopsis thaliana* [52,53]. The major aim of creating these populations was to generate lines displaying variation in heritable DNA methylation profiles, which can be used as a proxy for epigenetic variation. The two epiRIL

populations were created in the genetic background of the reference accession Columbia (Col-0), using lines carrying mutations in *METHYLTRANSFERASE 1* (MET1) or *DECREASE IN DNA METHYLATION 1* (DDM1). The mutants used were *met1-3* and *ddm1-2*, respectively. Loss of MET1 leads to widespread loss of DNA methylation and affects the redistribution of repressive histone marks [25,54] while loss of DDM1, a chromatin remodeler that is required for maintaining DNA-methylation, results in approximately 70% reduction in DNA methylation, particularly at transposable elements and repeats [26]. For both populations, individuals homozygous for wildtype DDM1 or MET1 in the segregating F2 were selected for inbreeding until the F8 (**Figure 1**). Curing of the mutant allele is essential to avoid the induction and accumulation of novel DNA methylation polymorphisms (alongside developmental defects) during the rounds of selfing [26,27,54]. Importantly, despite reestablishing the functional MET1 and DDM1 proteins, the DNA methylation polymorphisms that recombined in the F1 generation remain largely unaffected [52,53].

Stable parental epialleles (wild-type or mutant-derived) are inherited in a Mendelian fashion in the course of selfing, just as during normal RIL generation. This means that the frequency of heterozygosity (here ‘‘epi-heterozygosity’’) is declining by 50% in each subsequent epiRIL generation, resulting in <2% probability of epi-heterozygosity at any specific locus in the F7 (**Figure 1**). A population of epiRILs thus consists of individual lines each containing a mosaic of homozygous wild type and demethylated regions (DMRs) derived from the respective initial crossing parents (**Figure 1**).

A difference in developing the two mapping populations is that the *ddm1*-derived epiRILs went through one round of backcrossing the F1 to wildtype Col-0 (**Figure 1**), which reduced the amount of methylome divergence but at the same time allowed the creation of a large population of lines. The *met1*-derived epiRILs, lacking this round of

backcrossing, displayed unstable phenotypes and many could not be propagated by selfing due to accumulation of detrimental phenotypic effects or reduced fertility [53]. Interestingly, methylation polymorphisms for a few loci were still detected even in the F8 or F9 generation of *met1*-derived epiRILs in addition to *de novo* establishment of non-parental epialleles. This probably reflects cases of trans-chromosomal (de)methylation [79] and/ or progressive *de novo* methylation events [20]. Although it has yet to be determined to what extent such processes affect the currently available epiRILs, QTL^{epi} mapping procedures are challenged with this potential for non-Mendelian inheritance patterns (**Note 1**).

The available epiRILs show phenotypic variation in various traits, confirming the concept of DNA methylation affecting quantitative traits [52,55]. For specific purposes, it is possible to generate other sets of epiRILs in Arabidopsis following any of the explained strategies using different mutants that effect DNA methylation or any other epigenetic feature. However, the process is costly and time consuming and in order to be able to create a physical and/or genetic map of methylation markers it is necessary to perform extensive epigenotyping.

Retrieving epiRIL seed stocks

The approximately 500 *ddm1*-derived epiRILs are publicly available at the Arabidopsis Stock center of INRA Versailles (http://publiclines.versailles.inra.fr/). Furthermore, for subsets of those epiRILs whole genome methylation data [106] and DNA sequence data [56] is publicly available (123 and 52 lines, respectively).

The *met1*-derived epiRILs are available from their founders upon request [53].

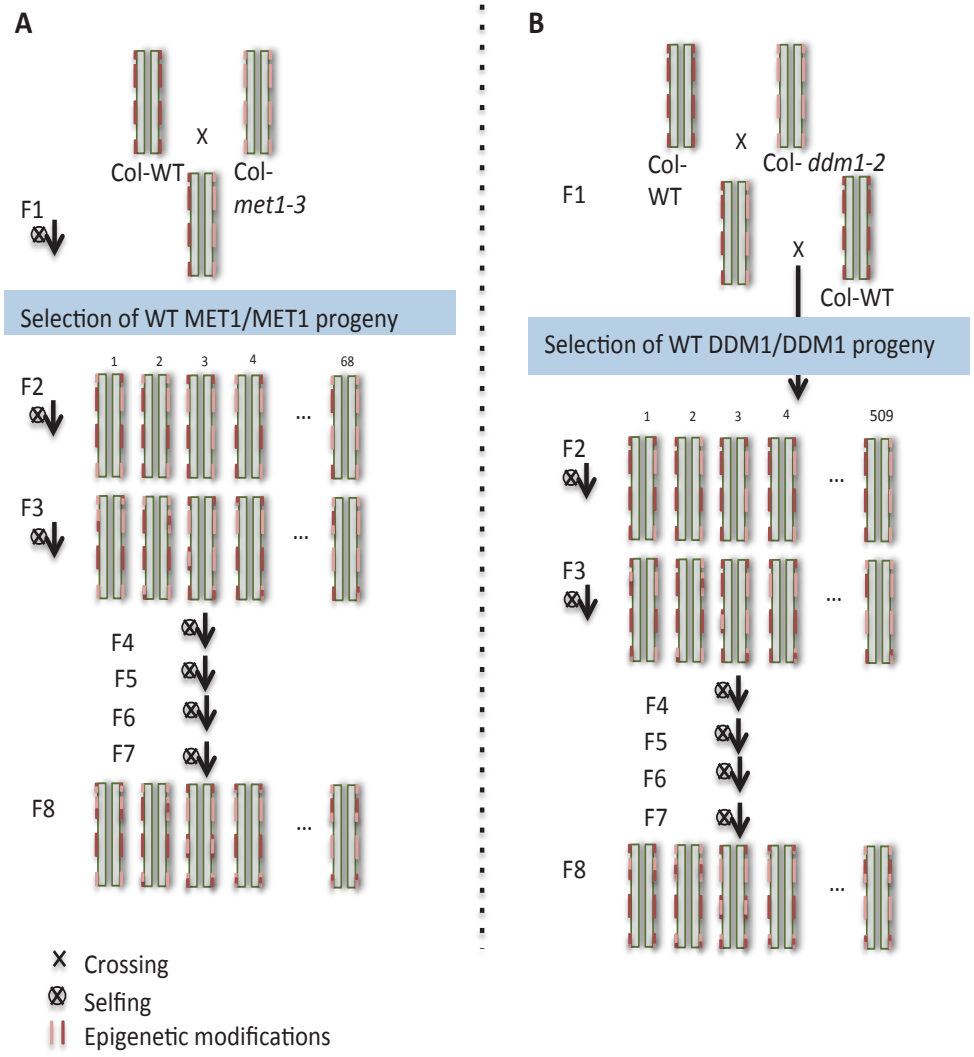

Figure 1: Construction of epiRILs. A) Construction of the *met1*–derived epiRILs: A Col-0 line was crossed to a Col-0 mutant line *met1-3* (MET1= METHYLTRANSFERASE 1) to create F1 hybrid plants. Selfed F1 plants gave rise to a segregating recombinant F2 population from which 96 individuals, homozygous for the wildtype MET1 locus, were selected for an additional six generations of selfing by single seed descent. Initially 96 epiLines were created but only 68 could

be advanced to the F8. By the F8 generation, lines reached (near) epi-homozygosity.

B) Construction of the *ddm1*-derived epiRILs: A Col-0 line was crossed to a Col-0 mutant line *ddm1-2* (DDM1= *DECREASE IN DNA METHYLATION 1*). The resulting F1 was backcrossed as female parent to a Col-0 line. Subsequently, about 500 progeny plants homozygous for the wildtype DDM1 allele were selected and propagated through six more rounds of selfing, generating a population of ~500 (near) epi-homozygous different epiRILs by the F8. Lines are depicted schematically as one chromosome.

3 Methods

To identify traits linked to heritable DNA methylation variation, a QTL^{epi} mapping analysis can be performed. QTL^{epi} mapping aims at identifying the epigenetic profiles at genomic loci that underlie a phenotypic trait of interest.

Before starting, it is important to carefully select the mapping population. Criteria that should be taken into account are a.o. a reasonable population size (which determines resolution and power of the mapping) and segregation of the trait of interest within the population. A good indication of the segregation of traits and the necessary population size can be obtained from phenotypic analyses of the population parents. Substantial trait variation might reflect large effect loci and suggests differential regulation in the parents, which will segregate in the epiRIL offspring.

3.1 Designing and Performing an Arabidopsis phenotyping experiment

1) Ensure stable conditions in the growth facility.
   In order to obtain phenotyping results that are consistent within and across experiments it is crucial to keep the growing environment as stable as possible [107] (**Notes 2**). In most facilities, conditions like

humidity, temperature and day/night cycle are standardly monitored and kept stable. Additionally, if possible, it is recommended to measure carbon dioxide levels and light intensity within the chamber and ensure that all plants receive the same amount of water and nutrients (see **Notes 3 & 4**).

2) To avoid so-called "seed batch effects", lines to be phenotyped should be propagated simultaneously under the same growing conditions (see **Note 5**) for at least one generation.

3) Homozygous lines allow including replicates of genetically identical lines. This provides a better estimate of the line specific trait value and cancels out random biological within-line-variation. Ideally, published data from the same lines can be used to estimate the level of variation in the population and to determine the number of replicates needed (power calculation).
   When designing experiments with limited numbers, e.g. when growth space is limited, experiments are expensive or many different treatments are compared, it is usually better to increase the size of the mapping population first before including replicates of identical lines. This will provide a higher resolution while the locus effect is replicated in different lines. Unless there are large segregation distortions, the allele frequency of each locus is centered around 50% in the population (**Note 6**).

4) Randomize the replicates of all lines, including the parents of the population and possibly their F1, throughout the growth/ greenhouse chamber (**Note 7**).

5) Stratify seeds for 3-5 days in the dark before sowing (**Note 8**).

6) Sow seeds and phenotype traits of interest in a systematic manner. For Arabidopsis that can be for example physiological traits like biomass, photosynthesis, or time to flowering or molecular traits like gene transcription, metabolomics or proteomics. However, every other trait that can be accurately quantified and which segregates in the population is admissible for QTL^{epi} mapping.

3.2 Data analysis

Descriptive Statistics

Quantitative traits display a continuous distribution of trait values, measured in different genotypes studied. Therefore, parameters of descriptive statistics, i.e. moments of trait value distributions, are routinely determined to summarize features of phenotypic datasets.

a) Parental variation, population mean and (coefficient of) variance

To estimate whether a trait will segregate in the mapping population the founding parents of the population can be analyzed. Such analyses can precede population measurements but it is advised to analyze the parents in the same experimental settings as the derived population. Strong differences in trait values between the parents is indicative of epigenetic causal variation and might provide leads to the sign, strength and number of QTL^{epi} segregating in the progeny. However, a lack of trait variation between the parents does not necessarily mean an absence of relevant epigenetic modifications. Multiple complementary QTL^{epi} with opposite effect signs might cancel out each other's effect in parental lines but might lead to transgression in population individuals (see below). To obtain accurate estimates of parental trait values it is recommended to include sufficient replicates in the analyses, especially when replicates of

population individuals cannot be included in the experimental design. In the latter case, within line variation of the parents can then be used as a proxy to estimate heritability in the segregating population (see below).

The population mean (μ) is determined from all individuals of a certain mapping population, as opposed to the line mean, which describes only the trait value of an individual epigenotype of a population. The standard deviation (SD) describes the extent of trait value variation in the population, i.e. between-line variation, which can be quite large for quantitative traits in plants due to transgressive segregation of multiple $QTLs^{epi}$. The SD of replicate measurements of isogenic individuals provides an estimate of within-line variation independent of epigenetic variation, e.g. due to biological variation or measurement error. Both between- and within-line variation is used in the calculation of trait heritability (see below). The standard error of the mean (SEM), which is derived from the SD and the sample size, is often determined as an additional measure of variation and shows how well the determined mean represents the population.

The variation of trait values in a population can be well described with the SD. However, for comparing variation between different epigenotypes, across experiments or between different traits it is useful to calculate the coefficient of variation (CV). The CV is the ratio between the SD and the mean ($CV = SD/\mu$) and thus shows the trait variability in a population in relation to the population mean.

b) Skewness and kurtosis, transformation of data

Most QTL-analysis software applies parametric tests and, therefore, assumes normal distributions of trait values in the population. However, in specific cases the distribution of trait values can deviate substantially from normality. Although QTL analysis is quite robust against deviations from normality, especially for large populations, transformation of data might

considerably increase mapping power and reduce false positive QTL detection.

Skewness describes the symmetry of a distribution while kurtosis provides a measure for the tails of the distribution (heavy- vs light-tailed). The most common deviations from normal distributed data are bimodal, skewed and platykurtic or leptokurtic distributions. Skewness and kurtosis of distributions can be calculated, for example, with the R/moments package. Bimodal distributions are often caused by the segregation of a single large-effect QTL, while the level of kurtosis depends on the number of small-effect QTLs and the level of residual non-genetic biological variation. Skewness can be caused by biological restrictions on one side of the spectrum, e.g. flowering does usually not occur before a specific developmental stage is reached but can be considerably delayed. Also, the choice of the unit of measurement can cause skewed distributions, e.g. trait values expressed as percentages cannot exceed the lower and upper boundaries of 0% and 100%, respectively, and these extreme classes, therefore, often result in a spike in the data set. The latter deviation from normality can be improved by a probit transformation, whereas other non-normal distributions benefit from a LOG or SQRT transformation. Data sets with strong positive skewness and leptokurtic distributions are usually LOG-transformed, especially if trait values range several orders of magnitude. SQRT-transformations, on the other hand, might improve the normality of negatively skewed and platykurtic distributions.

c) Heritability and explained variance

In quantitative genetics, heritability (H^2) is a measure to assess how much of the total variance in trait values is caused by genetic factors. Heritability estimates can for example be used for predicting the phenotypic value of offspring derived from a given cross (i.e. which parents should be chosen for a mapping population) or to give indications which strategies to follow

in breeding applications to alter the trait of interest. In terms of (epi)QTL mapping the H^2 estimates are useful to properly interpret the proportion of heritable variation explained by all the (epi) QTLs. For a single (epi)QTL its contribution to the phenotypic variance in the trait is often expressed as explained variance. The latter is often provided in the output of QTL analysis software but can also be assessed by ANOVA.

The total variance (Vp) in a quantitative trait is usually described as genetic variance (Vg) plus environmental variance (Ve) and the variance derived from the interaction between genetic and environmental factors (Vge).

$$Vp = Vg + Ve + Vge$$

While Vg can be assigned to specific genetic loci, Ve is random and constitutes biological variation, experimental and measurement error. Vg can be estimated from between-line variation, whereas Ve is estimated from within-line variation. To determine the latter adequately multiple replicate measurements of isogenic population individuals are necessary, although this is not a strict necessity for QTL mapping. When no replicate measurements of epiRILs are available, the within-line variation of the parents of the population can be used as a proxy for Ve, assuming equal variance in the different epiRILs of the population, which is another prerequisite of parametric tests. Genetic variance can be further split into additive (variance from additive gene effects), dominant (variance from dominant gene action) and epistatic variance (variance from interaction between genes).

Two specific types of heritability can be estimated, Broad-sense heritability (H^2) and Narrow-sense heritability (h^2).

Broad-sense heritability (H^2)

In the broad sense, heritability is estimated using the total genetic variance (Vg) divided by the total phenotypic variance (Vp) of a population of genetically diverse lines.

$$H^2 = Vg/Vp$$

In an experimental setup broad-sense heritability estimates indicate how much of the observed phenotypic variation can be explained by genetic or in this case epigenetic factors.

Narrow-sense heritability (h^2)

In the narrow-sense, heritability is measured by the genetic variance due to additive effects (from all loci influencing the trait) divided by the total phenotypic variance.

$$h^2 = Va/Vp$$

Since dominance cannot be estimated in homozygous populations, such as epiRILs, broad-sense heritability is usually calculated.

d) Transgression

The difference in trait values between the parents of an epiRIL population can be a good indication of the involvement of epigenetic factors in the regulation of a trait. In general the majority of the epiRILs should display intermediate trait values relative to their parents, but extreme phenotypes (outside the parental values) can occur. These are termed transgressive phenotypes [108]. Transgression is observed frequently, particularly in intraspecific crosses and when many opposite-effect QTLs segregate in the population. The distribution of transgressive trait values typically displays a leptokurtic shape. Possible explanations for transgressive segregation

are the accumulation and action of complementary-effect genes but also epistatic interactions or overdominance may contribute. The transgression and parental values of a trait, therefore, provide meaningful information on the level of epigenetic regulation and the epigenetic architecture of that trait.

3.2.2. QTL^{epi} mapping

a) Creating an epigenetic Linkage Map

The first step to link phenotypic variation in an epiRIL population to corresponding epigenetic loci is the creation of a linkage map. In RILs, genomic polymorphisms like single nucleotide polymorphisms (SNPs) are used as physical and genetic markers to determine which genomic fragment is derived from which parent. In epiRILs, differentially methylated regions (DMRs) that are stably inherited over generations serve the same purpose. To examine which DMRs are stably inherited over generations it is necessary to perform genome-wide DNA methylation analysis on the founder parents, determine the DMRs present and compare that data with DMRs still detectable in the selected epiRILs. Genome-wide bisulfite sequencing or Methylated DNA immunoprecipitation-sequencing (MeDIP) can be used for that purpose. Upon generating the necessary methylome data it is possible to determine which region is derived from which parent. DMRs that are stably inherited can be used as physical and genetic markers to create a map using software packages like MAPmaker [109] or JOINMAP [110].

To perform QTL mapping accurately it is important that the marker density is sufficiently high and that the spacing between markers is consistent in order to detect all cross overs. The principle assumption is that each locus on the genome is in linkage disequilibrium with one or more markers. Previously published epiRIL linkage maps reported a coverage of > 80% of

the Arabidopsis genome and a marker spacing of approximately 3.45 centiMorgan (cM) [56], where one cM refers to one recombination event per 100 meiosis events. Recombination during meiotic cell division results in the decay of linkage disequilibrium and as such the resolution of the population and the number of markers needed depends on the number of meiotic events. The resolution of a mapping population depends on the crossing setup for its creation. For example an extra round of back-crossing, as in the *ddm1*-derived epiRIL population stabilizes wildtype introgressions but also results in more crossovers than in the *met1*-derived epiRIL population. As a guideline, in classical Arabidopsis RILs, 5 cM correspond to approximately 1 megabase (Mb) and recombination frequency is one to two cross overs per chromosome per meiosis. Taking fixation due to inbreeding into account this means effectively two recombination events per chromosome in (epi)RILs. In small mapping populations, typically less than 200 individuals, this means that a spacing of 5 cM between markers should be sufficient to generate a saturating map.

b) Mapping epigenetic quantitative trait loci (QTL^{epi})

Once an epigenetic linkage map has been created for all selected epiRILs, QTL^{epi} analysis can be performed similarly to normal QTL analysis (**Figure 2**).

Interval mapping

Commonly used software is MapQTL, R/qtl or QTL cartographer. For $QTLs^{epi}$ analyses in particular the scanone function in R/qtl has been used. For this, the phenotyping data for the trait of interest is needed from the selected epiRILs. The segregation of trait values between lines is then compared to the segregation of epigenetic markers by stepwise screening along their chromosomes in windows and association scores are plotted

on the epiRIL linkage maps. Most QTL mapping approaches are based on interval mapping [111] which equally can be used for $QTLs^{epi}$ mapping. Interval mapping tests for the presence of a QTL every 2 cM or less between neighboring markers. At each locus, the method calculates the LOD (logarithm of the odds) score, which indicates the probability of a QTL at this position. If the LOD score exceeds a specific significance threshold, e.g. determined by permutation, the presence of a QTL in this region is more likely to be caused by epigenetic effects than by random chance. Note that the LOD score is logarithmic, so a LOD score of three indicates that it is a thousand times more likely that linkage is due to an epigenetic effect than due to random chance.

Calculation LOD:

H1: presence of a QTL H0: absence of QTL

$$\text{LOD score} = \log10 (L(\text{data} \mid H1) / L(\text{data} \mid H0))$$

Significance thresholds

In order to account for the genome-wide QTL search, genome-wide significance of LOD scores is usually determined by permutation testing [112]. Basically, permutation testing compares the observed LOD scores with the maximum LOD score that could be obtained by chance with permuted data (**Note 9**). Ideally, LOD significance thresholds should correspond to a false positive rate below 5%, which is often in the range of LOD 2 – 5.

Confidence intervals

To determine the most plausible genomic location of the detected QTL, confidence or support intervals are frequently constructed. For this, LOD scores of flanking loci around the maximum LOD score position are

determined and if they surpass a certain threshold the genomic location of the markers is included in the interval. Usually, two-LOD support intervals, meaning loci that reach a LOD score of the maximum LOD minus two units, are included. Genes, or any other genetic factors, located within this support interval can be considered as candidate genes explaining the observed variation caused by the QTLs^{epi}.

Multiple QTL mapping

If there is more than one QTL segregating in the plant population, the mapping approach can be adjusted to account for the effect of each individual QTL. One option for an adjusted mapping approach is composite interval mapping (CIM) or multiple QTL Mapping (MQM) where multiple marker regression analysis is combined with interval mapping. Basically, this means that the marker that is most strongly linked to the detected QTL is assigned as a co-factor, which absorbs the variation introduced by that QTL, and thus reduces residual variation for tests at other marker positions and increases the power to detect additional QTLs.

Explained variance

The H^2 estimates are useful to properly interpret the proportion of heritable variation explained by all segregating QTLs^{epi}.

A (multiple) regression model considering the nearest linked methylation marker can be applied to get an estimate of the contribution of each of the detected QTLs^{epi} to the total heritable trait variation. If heritability is high but the detected QTLs together explain just a part of the heritable variation, i.e. missing heritability, this means that there are more QTLs segregating that did not pass the significance threshold, possibly because of small-effect size and limited power. In addition, epistasis of two or more QTLs, whether detected or not, might explain considerable proportions of the total heritability.

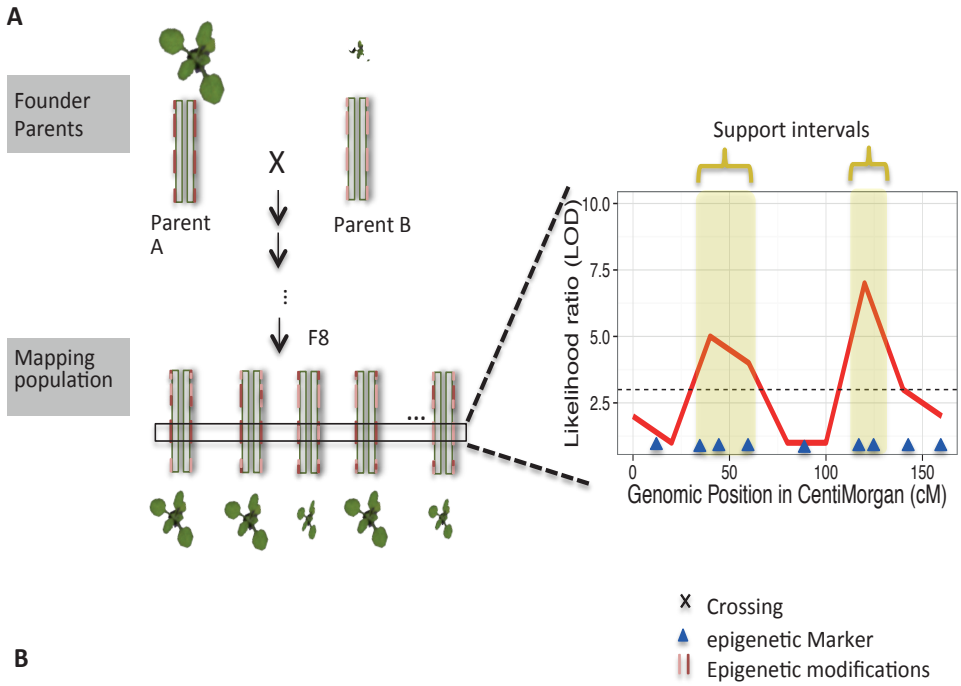

Figure 2: **QTL^{epi} mapping.** A) Schematic depiction of QTL^{epi} mapping. Two parental lines were chosen that are isogenic, except one has a mutation affecting the epigenome. The parental lines differ phenotypically in the trait for which QTLs^{epi} should be mapped (leaf area). The F8 generation is the mapping population. The parental lines and the mapping population are subjected to genome-wide analysis of stable epigenetic markers, which are in turn used to create a linkage

map. Phenotypic data on the trait of interest is generated and QTLs^{epi} are mapped by determining the LOD (logarithm of the odds) at all genomic intervals. The dotted line represents the threshold for a significant QTL^{epi} and the yellow shading indicates the support intervals around the peaks.
B) Stepwise approach of QTL^{epi} mapping

3.3 Follow-up analysis: confirmation, fine-mapping (cloning), validation and functional characterization

a) Confirmation of a QTL

One strategy to confirm the presence of a QTL is mendelising a QTL in an introgression line, i.e. a near isogenic line (NIL) harboring only the QTL interval of interest. A NIL is produced by several rounds of backcrossing a donor line (harboring the QTL) to a recurrent parental line, reducing the genomic introgression of the donor parent until only the QTL interval is present in an otherwise homogeneous recurrent background. After fixing the introgression by selfing, this line can be phenotyped and compared with the recurrent line. Principally, the same is possible for a QTL^{epi}, with the distinction that the DNA methylation profile of the introgression has to be traced by epigenotyping. With regard to the fact that transchromosomal (de) methylation events may happen at the introgressed fragment, additional genotyping for a genetic marker (if available) is recommended.

Steps to generate an epiNIL (**Figure 3**):

1. Cross an epiRIL containing the QTL introgression (and preferably as few as possible additional introgressions) with the recurrent wildtype parent.
2. Propagate the F1 by selfing to create a segregating F2.
3. Epigenotype the progeny for all possible segregating introgressions (**Note 10).**

4. Select the epiNIL, which harbors only the QTL introgression (**Note 11).**

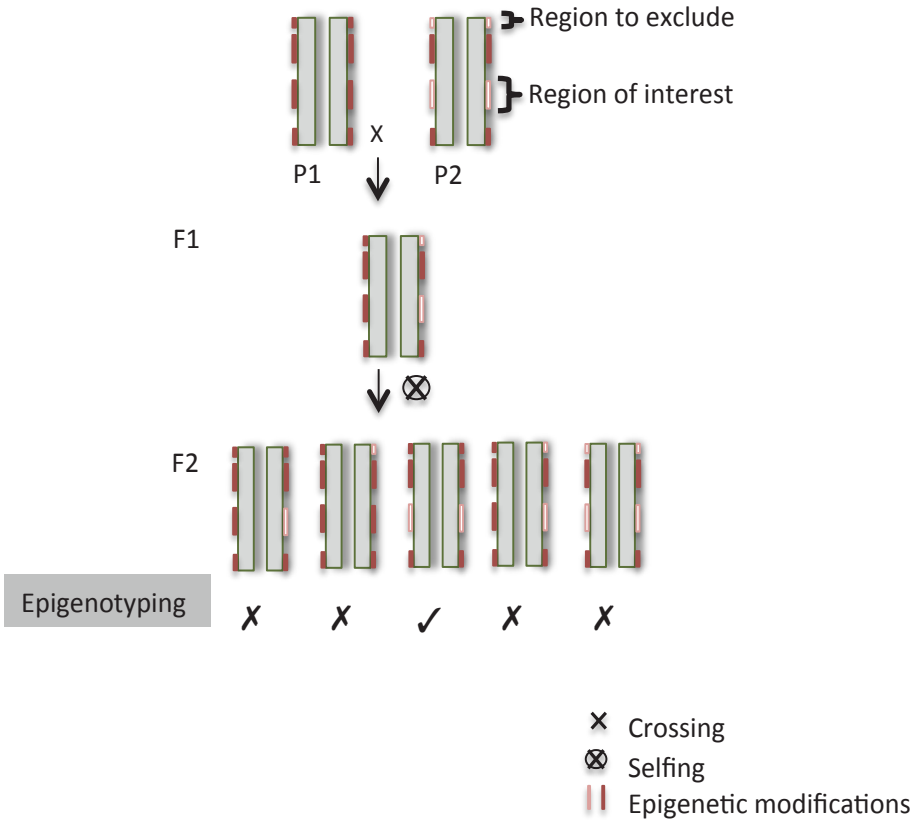

Figure 3: **Creation of an epigenetic near isogenic line (epiNIL) to confirm QTL^{epi} effects.** The recurrent parent (P1) is crossed to the epiRIL harboring the QTL region (P2). The F1 is selfed and the resulting F2 individuals are epigenotyped for the desired introgression. Epigenotyping should also be used to exclude introgressions unrelated to the QTL regions.

b) Epigenotyping Strategies

Two methods to determine methylation profiles at particular genomic positions are commonly used: endonuclease (McrBC) digestions followed by real-time PCR on marker positions and Targeted Bisulfite Sequencing. The first method is useful when a large quantity of (pooled) plants has to be assayed. This method is based on the endonuclease McrBC, which cleaves DNA containing methylated cytosines on one or both strands while it does not act on unmethylated DNA. Hence, a digest followed by quantitative (q)PCR on a marker region will allow quantifying the methylation at that position. To interpret the assay, controls need to be taken along that have the targeted position methylated and unmethylated (that can be for example the recurrent parents of the mapping population).

The second method, targeted bisulfite sequencing, is recommended if a better quantification of the methylation profile and/ or a higher resolution of a longer DNA stretch is desired. The method starts with bisulfite treatment of the genomic DNA, which results in conversion of unmethylated cytosines into uracil while methylated cytosines remain unaltered. Subsequently, the region of interest is amplified by PCR using unbiased degenerated primers (a primer mix that considers that each cytosine in the targeted region has the potential to be converted). The purified PCR product can either be sequenced directly or can be cloned into a vector with positive clones subjected to Sanger-sequencing.

Both methods establish the level of methylation of a specific genomic region and as such determine the descent of that region.

c) Fine-mapping and functional characterization

The next step, after identification and confirmation of an QTL^{epi} region, is determining the candidate genes in this region, which introduce trait variation through differential methylation. In addition, transposon insertions need to be excluded as causes for the QTL effect. Genome-

sequencing of a subset or the entire mapping population can reveal shared transposon insertions that could be causal. To date, genome-resequencing data is already publicly available for 73 of the *ddm1*-derived epiRILs [56]. Once transposon insertions in the QTL region have been identified in a subset of epiRILs, the remaining population can be screened by PCR for the same insertions.

Generally, QTL regions can be quite large (~ 1-2 Mb) and, therefore, may still contain hundreds of genes. Hence, it may be useful to assign *a priori* putative target genes for further study. Potential selection criteria for candidate genes are differential methylation between the recurrent parents or prior knowledge of gene functions that are related to the trait of interest.

Differential methylation patterns in the mapping population imply gene expression differences, which can be analyzed by quantitative Reverse transcription-PCR (RT-PCR). In Arabidopsis, both methylation at promoter regions but also gene body methylation can affect gene transcription [22,23,113] (**Note 12)**. To further confirm the causality of candidate genes targeted gene knock-downs or knock-outs can be analyzed. However, independent proof for the effect of DNA methylation at particular positions should be derived from cloning experiments, complementation or epigenetically modifying targets (region X methylated vs non-methylated).

Cloning of causal epigenomic loci is still challenging, as the DNA methylation profile might be altered during the process. However, methods to impose particular methylation profiles at loci of interest are available. For instance, transgenes can be used that induced the production of small RNAs, which in turn induce DNA methylation at their target site [114,115]. Also, techniques for epigenomic editing are rapidly emerging. These methods rely often on specific DNA recognition domains fused to a catalytic domain of a chromatin-modifying enzyme, allowing

targeting of the desired chromatin modification to any locus of interest [116]. Examples for such techniques are CRISPR-Cas, Zinc finger proteins or transcription-activator-like effectors (TALEs) [116].

3.4 Examples of Research Applications of epiRIL populations

epiRILs are a highly effective tool to assess the contribution of epigenetics to plant phenotypes. To date, they have been used to determine the contribution of epigenetic variation to quantitative traits [56], phenotypic plasticity in response to stress [57] or heterosis [58]. Excitingly, breeding strategies in canola crop plants already resulted in epiLines, which have improved energy use efficiency and drought resistance compared to their isogenic counterparts [59,60]. This is an example of artificial selection of certain epigenetic states, which enhance physiological plant characteristics. The next logical step here is defining the relevant genomic positions by approaches like QTL^{epi} mapping.

NOTES

1. One way to ensure that non-Mendelian inheritance patterns are not affecting mapping approaches is by thoroughly monitoring of epigenetic profiles of mapping populations (and propagated lines) either locus-specific or genome-wide and using only stable positions as markers.
2. Phenotyping can be done in both, growth or greenhouse chambers. However it should be realized that in the greenhouse environmental conditions are subject to more fluctuations and it is generally more difficult to keep a stable environment there.
3. Varying distances between plants and light source are common in growth chambers and can lead to plants being exposed to different light intensities. Basic (inexpensive) light sensors will allow to

determine strongly affected areas and to design an experimental setup avoiding them.

4. If no automated watering system is available, watering manually with the exact same amounts is recommended.
5. Different handling of the parent lines can cause seed batch effects [117]. If the parental lines of seed batches have been propagated in the same way this step is not necessary.
6. Strong epistatic effects involved in trait regulation increase the complexity in an experimental setup.
7. Randomization will level out phenotypic effects caused by plant position in the growth chamber.
8. Stratification will limit germination differences.
9. The number of permutations lies often between 1000 and 10,000.
10. In order to account for recombination it is wise to use one marker at each end of the introgression.
11. If the desired epiNIL is not identified perform another round of backcrossing or screen more F2 individuals.
12. For gene expression analysis it should be considered that fluctuations might occur across different tissue and also during the day (diurnal variation). Consequently, for comparative reasons it is important to profile the same tissue which was harvested at the same time of day from all individuals.

Funding
K.L. was supported by the Centre for Improving Plant Yield (CIPY), which is part of the Netherlands Genomics Initiative and the Netherlands Organization for Scientific Research.

Author contributions
K.L. and J.J.B Keurentjes wrote the manuscript.