Epigenetic and epigenomic variation in *Arabidopsis thaliana*  

Robert J. Schmitz¹,² and Joseph R. Ecker¹,²,³  

¹Plant Biology Laboratory, The Salk Institute for Biological Studies, La Jolla, CA 92037, USA  
²Genomic Analysis Laboratory, The Salk Institute for Biological Studies, La Jolla, CA 92037, USA  
³Howard Hughes Medical Institute, The Salk Institute for Biological Studies, 10010 North Torrey Pines Road, La Jolla, CA 92037, USA

*Arabidopsis thaliana* (*Arabidopsis*) is ideally suited for studies of natural phenotypic variation. This species has also provided an unparalleled experimental system to explore the mechanistic link between genetic and epigenetic variation, especially with regard to cytosine methylation. Using high-throughput sequencing methods, genotype to epigenotype to phenotype observations can now be extended to plant populations. We review the evidence for induced and spontaneous epigenetic variants that have been identified in *Arabidopsis* in the laboratory and discuss how these experimental observations could explain existing variation in the wild.

**Natural genetic variation**

Diversity within and between species is primarily driven by genetic variation and interactions with the environment. *Arabidopsis thaliana* has adapted to numerous environments, as it exists naturally and broadly throughout the northern hemisphere [1]. It is a selfing species that contains a largely isogenic genome, although residual heterozygosity does exist due to infrequent admixture [2]. The genome is compact and depleted of transposons and is relatively small (~150 Mb) compared to other plant species [3,4]. Recently, the *Arabidopsis* 1001 Genomes Project (see 1001genomes.org and signal.salk.edu/atg1001) was initiated with the goal of identifying the total genetic variation present within this species and to enable scientists around the globe to study their favorite quantitative trait of interest using association or linkage mapping techniques [1,5,6]. One challenge for adaptation of *Arabidopsis* to novel environments may be a limited set of mutations [7,8]. Therefore, it is essential to identify not only genetic variation but also natural epigenetic (see **Glossary**) variation, for which *Arabidopsis* is ideally suited [9,10]. In this regard, in addition to sequencing genomes [11–15], efforts are underway by the *Arabidopsis* 1001 Genomes Project (http://www.nsf.gov/awardsearch/showAward.do?AwardNumber=1122246) to sequence transcriptomes and DNA methylomes, which will be useful for identifying natural epigenetic variants.

**Epigenetic alleles (Epialleles)**

The extent to which epigenetic variation contributes to phenotypes remains to be determined [16]. However, several examples of phenotypic variation that can arise through epigenetic variation are known, including the *peloric* (*Linaria vulgaris*) [17], colorless non-ripening (*cnr* – *Solanum lycopersicum*) [18] and *B*’ (*Zea mays*) [19] epialleles. Studies in *Arabidopsis thaliana* have provided the richest source of known epialleles [20–32]. Many of the *Arabidopsis* epialleles require the activities of the RNA-directed DNA methylation (RdDM) machinery [33] for their formation and stability. RdDM functions to silence repetitive and transposon sequences and in some cases genes by a complex set of machinery that initiates transcription, generates 24nt RNAs and methylates targeted DNA in all cytosine contexts (for review see [33]). In fact, a number of these epialleles were used in genetic screens or as markers [21,22,28,29] to uncover novel components and mechanisms of the RdDM pathway [34–48].

Epigenetics is the study of mitotically and/or meiotically heritable changes in phenotype that arise independent of genetic variation. This definition is especially difficult to reconcile in studies of natural epigenetic variation, as by definition the populations under study exhibit extensive genetic variation. Recognizing this conundrum and the complexity of epiallele formation, a classification system that grouped epialleles into one of three classes based on their dependence on genotype has been proposed [49]. The first class called obligate epialleles display a complete dependency on a genetic variant, whereas the second class called facultated epialleles form from the presence of a

---

**Glossary**

**Accession:** a natural (wild) isolate from a specific geographic location.  
**CNV:** copy number variant.  
**DMR:** differentially methylated region.  
**Epiallele:** a heritable variant of an allele that is not reflected by a change in its DNA sequence, but does have an effect on gene expression.  
**Epigenetics:** the study of mitotically and/or meiotically heritable changes in phenotype that arise independent of genetic variation.  
**Epigenomics:** the genome-wide study of modifications to DNA and associated proteins such as a variety of modified and variant histones, nucleosome positioning, transcription factors as well as RNAs.  
**epiRIL population:** recombinant inbred lines that have wild-type genotypes, but contain mosaic epigenomes based on the parental lines used to establish the population.  
**MethylC-Seq:** a genome-wide technique to survey base resolution maps of DNA methylation using high-throughput sequencing.  
**RNA-directed DNA methylation (RdDM):** a mechanism that is involved in silencing through a feedback loop consisting of small RNAs guiding DNA methylation and histone modifications.  
**Small RNAs (smRNAs):** 21–24 nucleotides (nt) RNA species that can guide machinery to silence loci post-transcriptionally (21 nt smRNAs) and/or transcriptionally (24 nt smRNAs).  
**SMP:** single methylation polymorphism.
genetic variant such as a nearby transposon insertion, but their maintenance is not necessarily dependent on this variant. The third class, pure epialleles, form independent of any genetic variation. The epialleles thus far identified in Arabidopsis are either obligate epialleles [21,24] or are facilitated epialleles, as their formation has been induced using mutations [23,27,28,50] in machinery required for maintenance of DNA methylation or they were recovered from mutagenesis experiments [22,29]. Mitotically pure epialleles exist [51], but until recently [52–54] the identification of naturally occurring meiotically pure epialleles has remained elusive.

**Spontaneous epigenetic variation**

Epialleles have the potential for reversibility and often exist in metastable states, as has been observed for both the peloric and cnr natural epialleles [17,18]. With the exception of transposon-induced events, this feature is unique to epialleles as genetic mutations are typically considered unidirectional over evolutionary time. The molecular spectrum and rate of spontaneous mutations are relatively well understood especially when compared to our understanding of spontaneous epigenetic variation. Genome resequencing of a collection of Arabidopsis thaliana mutation accumulation (MA) lines, which were created from a lone founder by single seed descent growth for 30 generations [55], revealed just under one mutation per line per generation [56]. Recently, base pair resolution DNA methylomes of this same population were determined, which uncovered a rate of base level spontaneous variation in DNA methylation (single methylation polymorphisms – SMPs) that was at least four orders of magnitude greater than genetic mutations [52,53]. Furthermore, these studies captured the formation of pure epialleles, some of which resulted in significant transcriptional variation (~10- to 1000-fold change) of the affected locus [52,53]. One study proposed that these pure epialleles may have arisen through incomplete reinforcement of methylation states that occurs postfertilization [52] (Figure 1), because the sites of epiallelic variation overlapped significantly with loci that are also affected in the maintenance methyltransferase met1 and in the triple mutant of DNA demethylases rdd (REPRESSOR OF SILENCING1; DEMETER-LIKE 2; DEMETER-LIKE 3) [57–59]. The involvement of epigenetic mechanisms that require the activities of small RNAs and/or DNA methylation machinery are widespread in plant reproductive development. In fact, both the maternally and the paternally inherited genomes in Arabidopsis require intricate mechanisms for ensuring their identity and for inheritance of epigenomic information from generation to generation [60,61]. One feature of these mechanisms is to correct errors (loss of DNA methylation) through the use of smRNA (small RNA) guides [62], but inefficiencies in transferring this information across generations may result in the formation of a pure epiallele.

The variation in methylation observed in the MA lines was not uniformly distributed throughout the genome. This was especially the case for CG gene body methylation [9,63,64], which still has an unknown role and requires MET1 for its maintenance. In fact, there was a depletion of variation associated with transposon sequences, exons and the pericentromeres [52,53]. These genomic regions all are associated with nucleosome occupancy [65] and may reflect that methyltransferase enzymes such as MET1 maintain nucleosome-associated DNA methylation at a higher fidelity than nucleosome-free DNA methylation.

In general, the rates of spontaneous variation in DNA methylation were similar between all of the MA lines except for one (line 69). This hypervariability is unlikely due to the environment because the entire population was derived in a common greenhouse in synchrony, which would be expected to affect numerous lines if there were such pressures. Additionally, line 69 was identified as being hypervariable in two independent laboratories [52,53]. The basis for this hypervariability has been examined by resequencing the genome of line 69 [53]. Intriguingly, an additional nonsynonymous SNP (G→T, Ala→Ser) not previously reported in this line [56] was identified in MEE57 [53], a homolog of the maintenance methyltransferase MET1 [66]. It is conceivable that the variation in

---

**Figure 1.** Incorrect maintenance of methylation states over generational time at certain loci may lead to the formation of pure (nongenetic) epialleles. It has been proposed that the activities of MET1 and RDD are required to maintain methylation states in the germ line during postfertilization [52]. A failure to maintain the appropriate methylation state by plants wild type for either MET1 [(a) – loss of DNA methylation] or RDD [(b) – gain of methylation] or additional enzymes (X) may result in the formation of pure epialleles, unrelated to any cis or trans genetic variation. Gold lines, CG methylation; blue lines, CHG methylation; pink lines, CHH methylation.
DNA methylation observed in line 69 may be due to this mutation in MEE57, which could potentially represent one route to the formation of genome-wide epigenetic variation.

**Mutation-induced epigenetic variation**

As described above, spontaneous genetic variation may provide a source of natural epigenetic variation. Proof-of-concept for this type of epiallele formation was originally demonstrated by using mutants in *DDM1* (*DECREASE IN DNA METHYLATION 1* [67]), which encodes a SWI-SNF2-like protein that is related to chromatin remodelers in mouse and humans [68]. *ddm1* mutants result in a genome-wide depletion of DNA methylation, especially in repeated regions of the genome [67]. Remarkably, remethylation of *ddm1* chromosomes does not occur immediately upon introduction to wild-type chromosomes, instead remethylation occurs gradually over generational time [67]. Genome-wide depletion of DNA methylation [57,63,69] and epiallele formation [25] can also be induced in *met1* mutants, which can result in a variety of developmental phenotypes [70,71]. It is noteworthy that *ddm1* mutants do not display major developmental abnormalities until successive generations removed from the originally mutated plant [30,72]. However, much of this mutation-induced epigenetic variation results in the stable transmission of morphological phenotypes across generations, even when outcrossed from the original mutant backgrounds [23,27,50,73,74]. Although the major molecular phenotype of *ddm1* or *met1* mutants is a depletion of DNA methylation, examples of genetic variation in the form of genomic rearrangements, copy number variants (CNVs) and frequent DNA transposition have also been observed and may account for a considerable amount of phenotypic variability [21,50,73,75–77].

Barbara McClintock proposed the concept of gene regulation by transposon insertion within or near genes [78]. It is now commonly accepted that variation in gene expression can be regulated by the position of nearby transposons [79–82]. For example, transposon insertions in certain accessions of *Arabidopsis thaliana* can result in a rapid cycling flowering behavior [24]. Most fascinating is the recent identification of the ONSEN class of transposons that are heat responsive [83]. The transcriptional activation of ONSEN can persist for at least three days after the heat stress has been removed and can result in transposition in the germline of RNAi mutants [83]. Moreover, genome reinsertion of ONSEN in the RNAi mutants in the nonstressed next generation confers heat responsiveness to nearby genes [83]. Even though plant species have developed a wide range of mechanisms to ensure faithful transposon silencing, transpositions can occur over generational time either in the wild or in cell culture as well as in mutant backgrounds [62,76,83,84] and can provide a substrate for epigenetic variation to nearby genes (Figure 2).

**Epigenetic recombinant inbred lines (epiRILs)**

The relatively stable transmission of epigenetic variation observed in *ddm1* and *met1* mutants provided the basis for creation of epigenetic recombinant inbred lines or epiRILs [23,27]. RIL populations are commonly constructed in plants and are used for mapping causal variants for traits that are variable between two naturally occurring parental lineages. The epiRIL populations provide a unique opportunity to reduce genetic variability and increase epigenetic variability. Two independent epiRIL populations were created using either *ddm1* [23] or *met1* [27] mutants that were crossed to wild-type Col-0 plants to generate F1 hybrids. Segregating populations of plants were screened for the presence of wild-type alleles of either *DDM1* or *MET1* and propagated by single seed descent for multiple generations to increase homozygosity. Each epiRIL within the population essentially contains a mosaic epigenome derived from either wild type and *ddm1* or wild type and *met1*. Using microarray-based epigenomic approaches, it was determined that although some of the epialleles remained stable throughout construction of the epiRIL population, many slowly reverted back to the wild-type status over generational time [23,27]. This type of reversion is not uncommon and probably results from *de novo* methylation through smRNA guides [62]. In addition to epigenetic variation, there was clear evidence of genetic variation in the form of transpositions in the epiRILs [23,27,76].

Both epiRIL populations contained extensive phenotypic variation that was stable throughout each population for numerous traits such as time to flowering, plant height and fruit size [23,27]. Phenotypes present within the *met1* epiRIL population resembled variation observed in natural *Arabidopsis* accessions [27]. Moreover, significant heritability estimates and continuous distribution of phenotypes were observed in the *ddm1* epiRILs, indicating that multiple loci contribute to the observed phenotypic variation and these loci can potentially be identified through linkage.
mapping approaches [85]. Recently, the ddm1 epiRIL population was subjected to extensive phenotyping alongside natural Arabidopsis accessions in a common garden to test the contribution of epigenetic variation to phenotypic diversity [86]. The findings demonstrate that phenotypic variation in the epiRIL population displays heritability estimates similar to natural (wild) accessions grown in parallel [86]. Although epiRILs represent a set of artificially induced epialleles, these results provide compelling evidence that if spontaneous mutations do occur in components of the methylation machinery then an explosion of largely heritable epigenetic variation may quickly arise, providing a substrate for natural selection to act upon. As described above, a spontaneous mutation occurring in line 69 of the MA population was found in a component of methylation machinery (MEE57) [53], which could account for the increased hypervariability of methylation polymorphisms in these plants [52,53].

Natural epigenetic variation
Although natural epigenetic variation exists in Arabidopsis [9,10], understanding the basis for this variation presents a complex challenge due to the inherent genetic variation also present in these same accessions [11–15]. The experimental systems mentioned above provide alternative routes by which epigenetic variants can arise and spread within populations. There is, however, evidence for other events occurring in the wild [21,24,87,88]. A survey of Arabidopsis accessions for variation in methylation of the 178-bp centromeric satellite repeat identified hypomethylation of this region in the Bor-4 accession [88]. The causal variant for the hypomethylation was identified through linkage mapping as it segregated as a single recessive locus and encodes a methylcytosine binding protein named VIM1 (VARIANT IN METHYLATION 1) [88]. Because the methylation status of the centromeric repeat was entirely dependent on the presence of the VIM1 variant, methylation variation of centromeric repeats in Bor-4 represents an example of ‘obligate’ epigenetic variation [49]. This case nicely illustrates the ability of genetic variants to affect epigenetic variation in trans similar to what is observed in the epiRIL population. Another example of natural epigenetic variation that uncovers the complexity of epiallele formation exists at the PAI (PHOSPHORIBOSYLANTHRANILATE ISOMERASE) loci in certain Arabidopsis accessions [21,89]. In the Col-0 genome the PAI loci are present as a three-member gene family with high nucleotide sequence identity. In some accessions, the PAI1 locus exists as a duplicated inverted repeat as well as in other rearranged forms depending on the accession [89]. This inverted repeat leads to silencing by the RdDM pathway both in cis at PAI1 as well as in trans at PAI2 and PAI3 [45,90]. The methylated PAI1 loci provides an additional example of naturally occurring ‘obligate’ epialleles as the silencing of these loci dissipates when either PAI2 or PAI3 are segregated away by outcrossing from the inverted repeat at PAI1 [21].

Epigenomic approaches for studying natural epigenetic variation
The type of epiallelic variation present at the PAI gene family in some accessions of Arabidopsis is prevalent within this species. Using a comparative epigenomic approach, extensive methylation variation was identified between the Col-0 and Ler accessions; some of which overlapped with copy number variants identified within these same accessions [9]. Two major groups of methylation variants were identified in this study; one group contained hypervariable generic methylation patterns although a second group included transposons, which were invariably methylated. This second group was strongly associated with smRNAs typical of the RdDM pathway, whereas the genic methylation showed no such relationship [9].

An array-based epigenomic approach [10] was used to examine the genome-wide patterns and heritability of methylation polymorphisms between two different Arabidopsis accessions. This study revealed that inheritance of methylation was largely additive, as intermediate methylation levels were observed in F1 hybrids. They also determined that methylation variation immediately upstream of genes was strongly correlated with transcriptional variation, whereas variation of gene body methylation had no significant impact on gene expression levels [10].

There are currently a relatively small number of natural epialleles known in Arabidopsis, but with the use of epigenomic approaches this situation is likely to rapidly change [91]. Early attempts at profiling genome-wide patterns of histone modifications and DNA methylation resulted in low-resolution maps (~50–1000 bp) and were unable to provide information regarding specific (base resolution) DNA methylation contexts [9,10,64,92,93,63,94]. However, it is now possible to survey complete genomes for DNA methylation at base pair resolution by combining the gold standard for methylation detection sodium bisulfite conversion [95–101], with high-throughput DNA sequencing [57,69]. The cytosine DNA methylome maps produced using these new methodologies make it possible to distinguish between the commonly observed gene body CG methylation and the rarely observed gene body methylation that occurs in all three sequence contexts (CG, CHG, CHH), which is typically targeted by the RdDM pathway. As described above in the spontaneous epigenetic variation section, the ability to survey base resolution methylomes will rapidly accelerate the identification of natural epigenetic variants in any plant species that contains a reference genome.

Concluding remarks and future outlook
There exists natural epigenetic variation within Arabidopsis, but it is becoming increasingly clear that many forms of cis and trans genetic variation can contribute to the formation of epialleles. In the near future, it will be possible to use association mapping techniques that combine both nucleotide (CNV or SNP) and methylation variation (differentially methylated region (DMR) or SMP) present within this species to map causal variants for much of this epigenetic variation. It is also clear that perturbations to the machinery required for RNAi and/or DNA methylation probably result in a release of transcrip- tional silencing, promoting both genetic variation as well as the formation of novel epialleles. Methods now exist for detecting base resolution DNA methylomes, shifting the need for improvement to input sample preparation. Essentially all published studies of plant
epigenomes have utilized complex tissues, comprising numerous cell types, for DNA methylome profiling. Recent developments now enable enriching samples for specific cell types [102–104]. Future improvement may even allow interrogation of the epigenomes of single cells, which would provide unprecedented resolution for detection of epigenomic variation. These higher resolution views will be essential to unravel the intricate relationship between genotype, epigenotype and phenotype, allowing this knowledge to be applied to a wide range of species, including plants and people.

Acknowledgments

We would like to thank Miles Tanurdzic and anonymous reviewers for their helpful suggestions to improve the clarity and thoroughness of this review. R.J.S. was supported by an National Institutes of Health National Research Service Award (NIH NRSA) postdoctoral fellowship (F32-HG004830). This work was supported by the National Science Foundation (MCB-0929402 and MCB1122246), the Howard Hughes Medical Institute and the Gordon and Betty Moore Foundation (J.R.E.). J.R.E. is a Howard Hughes Medical Institute–Gordon and Betty Moore Foundation (HHMI–GBMF) Investigator.

References

19 Stam, M. et al. (2002) Differential chromatin structure within a tandem array 100 kb upstream of the maize b1 locus is associated with paramutation. Genes Dev. 16, 1901–1918
37 Chan, S.W. et al. (2006) RNAi, DRD1, and histone methylation actively target developmentally important non-CG DNA methylation in Arabidopsis. PLoS Biol. 5, e283

Singer, T. et al. (2001) Robertson’s Mutator transposons in A. thaliana are regulated by the chromatin-remodeling gene Decrease in DNA Methylation (DDM1). Genes Dev. 15, 591–602


Roux, F. et al. (2011) Genome-wide epigenetic perturbation jumps-steps patterns of heritable variation found in nature. Genetics 188, 1015–1017


Woo, H.R. et al. (2007) VIM1, a methylcytosine-binding protein required for centromeric heterochromatinization. Genes Dev. 21, 267–277


Shapiro, R. et al. (1973) Nucleic acid reactivity and conformation. II. Reaction of cytosine and uracil with sodium bisulfite. J. Biol. Chem. 248, 4060–4064


