



universität
wien

DISSERTATION / DOCTORAL THESIS

Titel der Dissertation / Title of the Doctoral Thesis

„Plant Phenomics and Epigenomics at Population Scale“

verfasst von / submitted by

Dipl.-Biochem. Patrick Hütter

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

Doctor of Philosophy (PhD)

Wien, 2022 / Vienna, 2022

Studienkennzahl lt. Studienblatt /
degree programme code as it appears on
the student record sheet:

UA 794 685 490

Studienrichtung lt. Studienblatt /
degree programme as it appears on
the student record sheet:

Molekulare Biologie

Betreut von / Supervisor:

Prof. Dr. rer. nat. Claude Becker

Acknowledgements

First and foremost, I would like to thank Claude Becker for his guidance and excellent mentorship over the last years. His talent to motivate, his patience and the freedom he gave me, enabled me to venture into new fields, learn plenty of new things and finally made me arrive somewhere, despite my fascination for potholes of all sizes that can be found along the road of science.

I would also like to thank Ortrun Mittelsten Scheid and Alexander Stark for agreeing to be on my thesis advisory committee and for providing helpful advice. Thanks to all past and present members of the Becker group, the VBC, and particularly the GMI for creating such a great and inspiring work atmosphere. I would like to thank Eva Rudolf, Jörg Hagmann and Christian Lindermayr for fruitful collaboration, Kelly Swarts for generously providing shelter in her group after the Becker group had left for Bavaria. Many thanks to the people involved in the VCDI for making the Vienna Biocenter an extraordinarily safe and mostly COVID-free place from the beginning of the pandemic. I would also like to thank Erich Birngruber for his guidance on all things HPC, and Borries Lubracki for his idea to put computer people into a separate room. I am also grateful to more people whom I was lucky to meet along the way, I sincerely thank Kathi, Niklas, Ranjith, Pierre, Ethan, Miguel, Ania, Falko, Lorenzo, Marco, Fabian and Moritz for countless hours of helpful discussion, inspiration, refreshment, and for being great comrades throughout this journey.

Last but certainly not least, I would like to thank my friends and family, who regardless of geographic distance always kept me grounded and attached to my roots. I am thankful to my parents for their constant support that enabled me to spend so many years at university in the first place. I am deeply grateful to my partner Silke for her never ending support and understanding, without you this work would have hardly been possible.

Abstract

Twenty years after the first drafts of the human genome were published, biology is going through a paradigm shift towards becoming a big data discipline. In plant biology, as in many other fields, the scientific community has accumulated a plethora of genomic resources that can readily be used to gain valuable insights about how the genetic makeup shapes the phenotype of an organism. Numerous studies have shown, however, that often the DNA sequence alone is not sufficient to fully explain the phenotypic outcome, and also that epigenetic mechanisms such as DNA methylation contribute to plant phenotypes. Furthermore, DNA methylation in plants is thought to bear adaptive potential by integrating environmental information into gene regulation, ultimately promoting phenotypic plasticity that allows plants to cope with ever-changing environments. Depending on these environmental factors, a single genome can thus give rise to many epigenomes, requiring further increase in data collection and population-scale analysis. The accumulation of large amounts of genomic and epigenomic data not only opens new avenues for scientific insight but also bears new challenges. With technological advances in DNA sequencing technology, both in terms of throughput and cost efficiency, the generation of whole-genome (epi)genomic information has become faster and cheaper. In contrast, acquiring phenotypic measurements at similar rate and scale remains time- and labor-intensive. For example, existing methods for image-based (semi-)automated plant phenotyping often have limitations regarding throughput or robustness towards changes in conditions during image acquisition. As such methods often rely on colour channel information, they are very sensitive to changes in illumination, plant colour, or background configuration. This thesis sets out to address how pretrained deep neural networks can be repurposed for fast and accurate high-throughput plant phenotyping and how such an approach, combined with identification of differentially methylated regions across populations, can lay important methodological groundwork for future work in population epigenetics.

Zusammenfassung

Zwanzig Jahre nach der Veröffentlichung der ersten Entwürfe des menschlichen Genoms hat die Biologie einen Paradigmenwechsel hin zu einer „Big Data“-Disziplin vollzogen. In der Pflanzenbiologie, wie auch in vielen anderen Bereichen, hat die wissenschaftliche Gemeinschaft eine Fülle von genomischen Ressourcen angesammelt, die herangezogen werden können, um wertvolle Erkenntnisse darüber zu gewinnen, wie der genetische Bauplan den Phänotyp eines Organismus prägt. Zahlreiche Studien haben jedoch gezeigt, dass die DNA-Sequenz allein den Phänotyp oft nicht hinreichend zu erklären vermag, sondern dass auch epigenetische Mechanismen wie die DNA-Methylierung zu Pflanzenphänotypen beitragen. Darüber hinaus geht man davon aus, dass die DNA-Methylierung in Pflanzen ein adaptives Potenzial beinhaltet, indem sie Umweltfaktoren in die Genregulation integriert und so letztlich die phänotypische Plastizität fördert, die es den Pflanzen ermöglicht, mit der sich ständig verändernden Umwelt zurechtzukommen. Abhängig von diesen Umweltfaktoren kann ein einziges Genom also multiple Epigenome hervorbringen, was letztlich eine weitere Steigerung der Datenerhebung, sowie Analysemethoden auf Populationsebene unabdingbar macht. Die Verfügbarkeit großer Mengen (epi)genomischer Daten eröffnet nicht nur neue Wege für wissenschaftliche Erkenntnisse, sondern birgt auch neue Herausforderungen. Mit technologischen Fortschritten in der DNA-Sequenzierung, sowohl im Hinblick auf den Durchsatz als auch auf die Kosteneffizienz, ist die Generierung ganzer (epi)genomischer Informationen schneller und kostengünstiger geworden. Im Gegensatz dazu ist die Erfassung phänotypischer Messungen in ähnlicher Geschwindigkeit und Größenordnung weiterhin zeit- und arbeitsintensiv. Bestehende Methoden zur bildbasierten, (halb-)automatisierten Phänotypisierung von Pflanzen haben zum Beispiel häufig Einschränkungen hinsichtlich des Durchsatzes oder der Robustheit gegenüber Veränderungen der Bedingungen während der Bildaufnahme. Da solche Methoden oft ausschließlich auf Farbkanalinformationen beruhen, reagieren sie sehr empfindlich auf Änderungen der Belichtung, der Pflanzenfarbe oder der Hintergrundkonfiguration. Diese Arbeit befasst sich mit dem Thema, wie vortrainierte tiefe neuronale Netze für die schnelle und genaue Hochdurchsatz-Phänotypisierung von Pflanzen eingesetzt werden können und wie ein solcher Ansatz in Kombination mit der Identifizierung differentiell methylierter Regionen in Pflanzenpopulationen wichtige methodische Grundlagen für zukünftige Forschung in der Populationsepigenetik schaffen kann.

Synopsis

This thesis sets out to narrow the persisting gap between phenotype and genotype. In Chapters 3 to 6, it aims to facilitate the study of epigenetic components of the phenotypic outcome, with a focus on DNA methylation, its genetic control, and the associated biochemical pathways that orchestrate its homeostasis. Chapter 2 presents a method to identify differentially methylated regions from whole genome bisulfite sequencing (WGBS) that is tailored specifically to plant epigenomes and is applicable on a population scale. One key result of Chapter 2, obtained from conducting large-scale genome wide association studies (GWAS) in the context of CHG methylation, indicates natural variation in a microRNA locus that contributes to homeostasis of CHG methylation. This microRNA and the post-transcriptional regulation it exerts on the dynamics of CHG methylation is the subject of Chapter 3. In Chapter 4, the methodology introduced in Chapter 2 is employed to study the epigenetic consequences of defects in a metabolic pathway that yields the central cellular methylgroup donor S-Adenosylmethionine (SAM). The last part of this thesis approaches the aforementioned gap from the opposite direction. It thematises the measurement of phenotypic traits and aims to address the current imbalance between the astounding throughput of current genomics technologies and the state-of-the-art for accurate quantification of plant phenotypes from large datasets. To that end, Chapter 5 presents a scalable and automated workflow for extraction of phenotypic measurements from image data.

Contents

Acknowledgements	i
Abstract	iii
Zusammenfassung	v
Synopsis	vii
Contents	ix
1 Introduction	1
1.1 Epigenetics	2
1.1.1 Chromatin Structure	2
1.1.2 Histone Modifications	3
1.1.3 DNA Methylation	4
1.1.4 Methods for Assessment of Genome-Wide DNA Methylation	5
1.2 High-Throughput Plant Phenotyping	8
1.2.1 Methods in Plant Phenotyping	9
1.2.2 Deep Learning in Plant Phenotyping	9
2 MethylScore, a Pipeline for Accurate and Context-Aware Identification of Differentially Methylated Regions from Population-Scale Plant WGBS Data	15
2.1 Abstract	16
2.2 Introduction	17
2.2.1 Features of Plant DNA Methylation	17
2.2.2 Whole-genome DNA Methylation Analysis as a Means to Understand Natural Variation and Stress Response	18
2.2.3 The Challenges in Determining Differential DNA Methylation . . .	18
2.3 Results	20
2.3.1 Outline of the Modular MethylScore Pipeline to Call DMRs from Plant WGBS Data	20
2.3.2 Segmentation of Plant Genomes with Different DNA Methylation Composition	22
2.3.3 MethylScore Identifies Differentially Methylated Regions without <i>a priori</i> Sample Information	22
2.3.4 Clustering of Methylated Regions Across Regenerated <i>A. thaliana</i> Populations Underlines Partial Maintenance of Organ-Specific Methylation Profiles in CG	24
2.3.5 DMR Calling on a Population Scale can Identify Unknown Underlying Sample Structures	26

2.3.6	MethylScore DMRs Reveal Recurrent <i>Trans</i> -Acting Association Signals in <i>A. thaliana</i> Natural Accessions	28
2.4	Discussion	30
2.4.1	MethylScore does not Require <i>a priori</i> Information on DNA Methylation Parameters or Dataset Structures	31
2.4.2	A Modular Pipeline That can be Integrated With Other WGBS Analysis Pipelines	31
2.4.3	Limitations and Shortcomings of the Pipeline	32
2.5	Outlook	32
2.6	Materials and Methods	32
2.6.1	Pipeline Strategy	32
2.6.2	Implementation in Nextflow	33
2.6.3	Re-analysis of Published Datasets	33
2.6.4	1001 Methylomes Analysis	33
2.6.5	Data Visualization	34
2.7	Acknowledgements	34
2.8	Author Contributions	35
2.9	Competing Interests	35
2.10	Data Availability	36
2.11	References	36
2.12	Supplement	43
3	Repression of CHROMOMETHYLASE 3 Prevents Epigenetic Collateral Damage in Arabidopsis	49
3.1	Abstract	50
3.2	Introduction	50
3.3	Results	52
3.3.1	Cell Division is Linked with CG and CHG Methylation Through Distinct Mechanisms	52
3.3.2	Genome-wide Coordination of Symmetric DNA Methylation	52
3.3.3	Repression of CMT3 During Embryogenesis Regulates Methylome Dynamics	55
3.3.4	Chromatin Features Associated With CMT3-induced Gene Methylation	57
3.3.5	Impact of CMT3-Induced Hypermethylation on Gene Expression	59
3.4	Discussion	60
3.5	Materials and Methods	64
3.5.1	Plant Material and Growth Conditions	64
3.5.2	Generation of Transgenic Lines	64
3.5.3	Generation of CRISPR/Cas9 Knockout Mutants for MIR823	65
3.5.4	qRT-PCR Analysis	65
3.5.5	Sample Size Estimation, Embryo Isolation and Nucleic Acid Extraction	65
3.5.6	DNA Methylation Profiling and Analysis	65
3.5.7	mRNA Profiling and Analysis	66
3.5.8	ChIP-Seq Analysis	66
3.5.9	Metaplots	66
3.5.10	CMT3 Transgene Copy Number Estimation	67
3.6	Acknowledgements	67

3.7	Additional Information	67
3.7.1	Author Contributions	67
3.7.2	Author ORCIDs	67
3.7.3	Decision Letter and Author Response	68
3.8	Additional Files	68
3.8.1	Supplementary Files	68
3.8.2	Data Availability	68
3.9	References	69
4	GSNOR Contributes to Demethylation and Expression of Transposable Elements and Stress-Responsive Genes	75
4.1	Abstract	76
4.2	Introduction	76
4.3	Materials And Methods	78
4.3.1	Plant Material and Cultivation	78
4.3.2	Quantification of MTA, SAM, SAH, Cys GSH, and Hcys	79
4.3.3	Whole-Genome Bisulfite Sequencing and Data Analysis	79
4.3.4	RNA Sequencing	80
4.3.5	Acid Extraction of Histones	80
4.3.6	Accession Numbers	81
4.4	Results	81
4.4.1	GSNOR and SAHH1 are Involved in Regulating Metabolite Levels of the Methylation Cycle	81
4.4.2	Loss of GSNOR1 and SAHH1 Functions Results in Altered Histone Methylation Levels	82
4.4.3	SAHH1 and GSNOR1 Functions Affect DNA Methylation	84
4.4.4	GSNOR1 and SAHH1 Regulate DNA Methylation of TEs and Genes	86
4.4.5	Transcriptomic Profiling of <i>gsnor1-3</i> and <i>sahh1</i> Plants	89
4.4.6	Integrative Analysis of WGBS and RNA-Seq Data	91
4.5	Discussion	93
4.5.1	Loss of GSNOR1 Function Results in an Increased Methylation Index	93
4.5.2	GSNOR1 Function is Crucial for the Maintenance of Histone Methylation and DNA Methylation	94
4.5.3	Alteration in DNA Methylation does not Correlate with Gene Expression	95
4.5.4	GSNOR1 Regulates Demethylation and Expression of TEs and Stress-Responsive Genes	95
4.6	Conclusions	97
4.7	Supplementary Materials	97
4.8	Author Contributions	98
4.9	Acknowledgements	98
4.10	References	98
4.11	Supplement	104
5	ARADEEPOPSIS, an Automated Workflow for Top-View Plant Phenomics using Semantic Segmentation of Leaf States	115
5.1	Abstract	116

5.2	Introduction	116
5.2.1	The Phenotyping Bottleneck in Plant -Omics Studies	116
5.2.2	Challenges in Image Analysis	116
5.2.3	Machine Learning Methods in Plant Phenotyping	117
5.3	Results	118
5.3.1	Color-Guided Segmentation can be Misguided by Image and Object Parameters	118
5.3.2	Training of a Proof-Of-Principle Model	119
5.3.3	Advanced Models for Differentiated Plant Area Classification of Individual Leaves	119
5.3.4	Extraction of Morphometric and Color Index Traits from Segmented Images	121
5.3.5	The ARADEEPOPSIS pipeline	122
5.3.6	Validation and Generalization to Public Phenotyping Datasets and Comparison to other Models	122
5.3.7	Validation of Color Index Traits	124
5.3.8	Validation of Pixel-Level Classifications	125
5.3.9	Analysis of other Plant Species	126
5.4	Discussion	126
5.4.1	A Fast and Easy-to-Use Segmentation Tool	126
5.4.2	Accurate Determination of Plant Health State	127
5.4.3	Limitations and Shortcomings of the Proposed Models	127
5.5	Outlook and Perspective	127
5.6	Methods	128
5.6.1	Plant Growth and Image Acquisition	128
5.6.2	Training and Validation	128
5.6.3	Implementation	128
5.6.4	GWA Studies	128
5.7	Acknowledgements	129
5.8	References	129
5.9	Supplement	131
6	Discussion	137
	References	139
	List of Acronyms	151

1 Introduction

Since the dawn of the post-genomic era, the availability of genomic resources has grown in abundance and continues to do so. With the rise of *omics* technologies, parts of modern biology have started to transition towards becoming a *Big Data* discipline, thereby shifting paradigms from classical gene-by-gene approaches to investigate one protein at a time, towards more top-down strategies that leverage and combine multi-omics data to gain scientific insight. Since the year 2000, when the genome of the model plant *Arabidopsis thaliana* was published (Arabidopsis Genome Initiative 2000), hundreds of plant genomes have been sequenced and there are ongoing efforts to extend that number by a large margin within a few years (Cheng et al. 2018). In the model organism *Arabidopsis thaliana*, long-standing collaborative efforts have led to an enormous genomic resource including extensive transcriptomic, genomic, and epigenomic information across geographically diverse habitats of the same species (Kawakatsu et al. 2016; 1001 Genomes Consortium 2016). A key goal of basic plant research is to understand how the genotype of an organism gives rise to its phenotype and its often remarkable plasticity. Soon after genomes had been sequenced, researchers started to build catalogues of single-nucleotide polymorphisms aiming to use them in genome wide association studies (GWAS) in order to find associations between the occurrence of a given mutation with a phenotype, such as disease or plant growth. While prospects were high, population geneticists soon became aware of the *missing heritability problem* (Manolio et al. 2009), revealing that the effects of single genetic variants on phenotypic outcome had long been overestimated. Part of this missing heritability lies in the polygenic nature of complex traits, often leading to only marginal contributions of single variants to a trait. Another part is caused by the environmental influence on the phenotype, the mechanisms that enable organisms to perceive and integrate environmental cues have yet to be elucidated in detail. Some plant species can display stunning phenotypic diversity even in near-isogenic populations, highlighting the potential of epigenetic contributions towards shaping the phenotype, which has become evident from the discovery of phenotypically relevant epi-alleles (Cubas et al. 1999; Ong-Abdullah et al. 2015; Miura et al. 2009).

1.1 Epigenetics

The term epigenetics has its origins in the field of embryology (Waddington 1942), where it was initially coined to describe the mechanisms of development that, starting from a single fertilised egg with its inherited information - later known as the genome - lead to the formation of different cell types and thus give rise to complex tissue structures and the entire body plan in a multicellular organism. Back then, the concept of *epigenesis* stood contrary to still debated *preformationist* views, according to which a fetus develops by mere growth of a *homunculus*, a preformed entity that is already contained within the zygote. Much has changed since then, the sequential processes underlying epigenesis are now referred to as cell differentiation, but at the time the fundamental gap between genotype and phenotype was already well perceived. Remarkably so, in light of the fact that this was before DNA had even been identified as the carrier of genetic information (Avery et al. 1944; Hershey and Chase 1952), its structure resolved (Watson and Crick 1953) and its mode of replication understood (Meselson and Stahl 1958). The definition of epigenetics has evolved since then. After the sequential nature of heritable information in the form of DNA was understood, epigenetics was redefined to describe mechanisms auxiliary to transcription of DNA, determining which information is expressed in particular cells, ultimately resulting in phenotypic differences (Nanney 1958). The following decades shone more light on these mechanisms, after chromatin structure was characterised (Kornberg and Thomas 1974), and the roles of histone modifications (Allfrey et al. 1964) and DNA methylation (Holliday and Pugh 1975; Riggs 1975) in gene regulation were identified. The discovery of stable transmission of these covalent modifications throughout mitotic cell divisions (Bird and Southern 1978), and the heritable phenomenon of paramutation (Brink 1956), that was later shown to be caused by alterations in DNA methylation patterns (Stam et al. 2002), led to a repeated redefinition of epigenetics to account for its heritable aspects of gene regulation that are not explainable by the changes in DNA sequence (Russo et al. 1996). There is still ongoing debate about whether heritability is a required aspect in the definition of epigenetics, because some of the known modifications appear to be transient and have not been shown to be transmitted meiotically. Whether heritable or not, the following sections of this Chapter shall focus on epigenetic mechanisms within the scope of plant biology.

1.1.1 Chromatin Structure

Within eukaryotic cells, DNA is not present in an open conformation but instead is tightly associated with proteins to form the structure of chromatin that readily fits into the spatially constrained nucleus. This structure serves to compact the genetic information by many orders of magnitude and at the same time coordinates vital cellular functions, including DNA repair (Osley et al. 2007), faithful replication of heritable information for cell division (MacAlpine and Almouzni 2013), and the regulation of gene expression (Venkatesh and Workman 2015). Mechanistically, this is realised by a combination of different hierarchical layers of compaction that ultimately result in the formation of chromosomes. The basal layer is achieved by nucleosomes, spheroid octameric protein complexes (Thomas and Kornberg 1975) that are formed by heterotetrameric assembly of core histones, combining two H2A-H2B dimers and a tetramer comprised of two molecules of histones H3 and H4, respectively. DNA is repetitively wrapped around these structures,

with each nucleosome core unit accomodating 146 bp of DNA (Luger et al. 1997) and is interspersed by nucleosome-free linker regions. This organisation forms a structure that resembles beads-on-a-string (Olins and Olins 1974). A second layer of compaction is achieved in presence of histone H1 that binds the linker regions, thus inducing further coiling of the former into a higher-order helical structure, the so-called 30-nanometer fiber (Bednar et al. 1998). Due to the resulting differences in DNA accessibility, these layers of compaction contribute to distinct regulatory domains that are generally referred to as transcriptionally active and gene-rich euchromatin, and transcriptionally silent heterochromatin, respectively. Chromatin structure and the regulation it exerts on gene expression is highly dynamic and depends on the exact mechanisms that act on its remodeling, including covalent histone modifications that act to define chromatin states.

1.1.2 Histone Modifications

Structurally, each nucleosome subunit contains a globular domain, the evolutionarily conserved histone fold, which is accompanied by an N-Terminal region rich in basic residues that is protruding outward from the histone core. These histone tail regions contain sites of extensive post-translational modification, including phosphorylation (Nowak and Corces 2004), (mono-/di-/tri-) methylation (Zhang and Reinberg 2001), acetylation (Sternier and Berger 2000), and ubiquitinylation (Shilatifard 2006), amongst others (Hagiwara et al. 2002; Cuthbert et al. 2004; Hassa et al. 2006; Nathan et al. 2006; Sabari et al. 2016). Among the first post-translational modifications of histones that were discovered is the acetylation of ϵ -amino groups in lysine side chains (Allfrey et al. 1964), which is regulated by a highly conserved set of antagonistically acting enzymes, histone acetyl transferases (HATs) and histone deacetylases (HDACs) (Kurdistani and Grunstein 2003). Histone acetylation was initially solely attributed to transcriptional activation based on the stabilising electrostatic interaction between the negatively charged phosphodiester backbone of DNA and the positively charged residues in histone tail regions. Transfer of acetyl groups to lysine residues mask the positive charge of the amino group and thus weaken such interactions and in consequence facilitate access to DNA by the transcriptional machinery. The study of histone modifications has been extended and changed over time, as emerging evidence has given rise to the histone code hypothesis (Strahl and Allis 2000; Jenuwein and Allis 2001; Kouzarides 2007), highlighting the interplay between different histone modifications as well as the modification of nucleosomes through dynamic incorporation of histone variants, homologous proteins with sequence divergence and distinct functional roles during development (Jang et al. 2015), cell division (Sullivan et al. 1994), DNA damage response (Morrison and Shen 2005), or chromatin compaction (Wang et al. 2018). A central component of the histone code lies in the existence of reader domains such as bromo- (Owen et al. 2000; Filippakopoulos et al. 2012), chromo- (Lachner et al. 2001; Jacobs and Khorasanizadeh 2002; Nielsen et al. 2002), or tudor domains (Lu and Wang 2013) that show binding specificity towards certain histone modifications. Once recruited, these domains directly aid the deposition of epigenetic marks or take scaffolding roles in the downstream recruitment of different histone modifiers, leading to formation of affinity binding sites for distinct reader or eraser domains. By this mechanism, regulation of gene expression is imposed on the epigenetic level by a repertoire of modifications that are subject to highly dynamic modulation with substantial crosstalk between different modifications (Fischle et al. 2003). A wide variety of covalent histone modifications have been characterised

to date. Depending on the underlying mechanistic, they possess either repressive or permissive characteristics with regard to the transcriptional output of the genomic regions they occupy. Euchromatic marks such as methylated H3K4, acetylated H3K9 or H3K27 are typically associated with actively transcribed genes (Litt et al. 2001; Grunstein 1998), whereas silent heterochromatic regions are mainly characterised by hypoacetylation and presence of H3K9 methylation (Richards and Elgin 2002). Particularly in the context of heterochromatin, there is not only substantial crosstalk between histone modifications but also with DNA methylation, a distinct epigenetic mechanisms that shall be introduced in the next section.

1.1.3 DNA Methylation

DNA methylation is a well-studied epigenetic mark found in plants and animals alike (Holliday and Pugh 1975; Bird 2002). Although there are different modifications known that involve the transfer of a methyl group to nucleobases present in DNA and RNA (Dunn and Smith 1958; Ehrlich et al. 1985), the term DNA methylation as it is used here shall refer to the addition of a methyl group to position five of the pyrimidine ring in cytosine bases (Figure 1.1). While many of the epigenetic mechanisms described before show tremendous conservation in most eukaryotes, others seem to have diverged in a more lineage-specific manner. In plants, DNA methylation occurs in three different DNA sequence contexts (Gruenbaum et al. 1981; Meyer et al. 1994): symmetric CG and CHG methylation as well as asymmetric CHH methylation (where H represents any nucleobase but G). This is different from DNA methylation in animals, where CG methylation is the predominant type (Goll and Bestor 2005; Bernstein et al. 2007). Symmetry in this case refers to methylated cytosines being present on both DNA strands through base complementarity to guanine and has implications on the transmission of such epigenetic information, as DNA replication during cell division necessitates the (re-)establishment of methylation on daughter strands. In the case of symmetric methylation contexts, the methylated state persists throughout semi-conservative DNA replication on one strand, thereby allowing maintenance mechanisms to faithfully retain this epigenetic modification by recognition of hemimethylated template DNA. In the CG sequence context, DNA METHYLTRANSFERASE 1 (MET1) is recruited together with VARIANT IN METHYLATION 1-3 (VIM1-3) to transfer methyl groups to CG sites on the opposite strand of hemimethylated DNA (Woo et al. 2008). Maintenance of CHG methylation has been shown to be interdependent with the histone tail modification H3K9me2, which is targeted by CHROMOMETHYLASE 3 (CMT3) via its chromodomain to specifically methylate CHG sites (Lindroth et al. 2001), which in turn can reinitiate H3K9 deposition via the histone methyltransferases KRYPTONITE (KYP) and SUPPRESSOR OF VARIATION 3-9 HOMOLOG (SUVH4/5/6) proteins, thus establishing a self-reinforcing loop (Jackson et al. 2002; Du et al. 2014). For asymmetric CHH methylation on the contrary, the methylated state needs to be (re)-established *de novo* by specific targeting of respective regions by CHROMOMETHYLASE 2 (CMT2), that is recruited to methylated histone tails via its chromodomain to methylate adjacent CHH sites, as well as by DOMAINS REARRANGED METHYLTRANSFERASE 2 (DRM2). The latter is a component of RNA-directed DNA methylation (RdDM) (Wassenegger et al. 1994; Cao et al. 2003), a silencing mechanism present in flowering plants that acts to repress transposon activity by virtue of the plant-specific RNA-Polymerases Pol IV and Pol V giving rise to RNA transcripts. These small RNAs are subject to processing by DICER-LIKE (DCL)

endonucleases and subsequently loaded onto ARGONAUT proteins, allowing for specific targeting of genomic sequences with complementarity to the resulting 24-nt siRNAs by recruitment of *de novo* methyltransferases such as DRM2. Importantly, RdDM acts not only on CHH sites but also in other sequence contexts (Law and Jacobsen 2010). These distinct molecular machineries controlling DNA methylation fulfill distinct biological roles, depending on the genomic contexts or chromatin environment they occur in. In *A. thaliana*, only a small fraction of the genome is methylated. This methylation is unevenly distributed, both in abundance and the genomic context it occurs in. CG methylation, measured as the percentage of methylated sites given the total number of CG positions, occurs most frequently (ca. 24%) and is associated with genic regions as well as heterochromatic regions rich in repetitive sequences and transposable elements (TEs). In contrast, CHG (ca. 7%) and CHH (ca. 2%) methylation is primarily localised in heterochromatic regions (Cokus et al. 2008; Lister et al. 2008). Functionally, one key role attributed to DNA methylation is the silencing of transposons to prevent their deleterious effect to genome integrity (Zhang et al. 2006). RNA transcripts involved in RdDM target repetitive regions of the genome such as TEs, typically resulting in relatively higher CHH methylation compared to other genomic regions. In contrast, gene body methylation (gbM) occurs mainly in CG context and is thought to be a signature mark of actively transcribed genes, yet its biological significance and functional implications are debated (Bewick and Schmitz 2017; Takuno et al. 2016).

1.1.4 Methods for Assessment of Genome-Wide DNA Methylation

Whole Genome Bisulfite Sequencing

DNA methylation can be measured on the genome-wide scale using next-generation sequencing techniques. Whole genome bisulfite sequencing (WGBS) on the Illumina[®] platform has been considered the gold standard for assessment of genome-wide DNA methylation for nearly two decades, although it is not unlikely to be superseded by more recent advances in long-read sequencing technology capable of detecting base modifications from native DNA material (Simpson et al. 2017).

In an endothermic, acid-catalysed reaction, non-methylated cytosine is deaminated to uracil in presence of sodium bisulfite via an intermediate sulfonyl adduct (Figure 1.1). Conversion of 5-methyl-cytosine however is kinetically inhibited and thus occurs at a negligible rate. Subsequent PCR amplification of bisulfite-treated DNA with a uracil-tolerating polymerase during library preparation enables the read-out of former cytosine positions as thymine positions, whereas methylated cytosines will still be read out as cytosine positions (Frommer et al. 1992).

Alignment of Bisulfite-Converted Sequencing Reads

Accurate mapping of reads obtained from sequencing bisulfite-converted DNA is challenging because of the resulting ambiguity of non-methylated cytosine and thymine positions. Different approaches have been used to overcome this problem, albeit with increased computational demand compared to standard reference-based alignment. The most common approach is the alignment against a three-letter genome as implemented by **bismark** (Krueger and Andrews 2011) or **bwameth** (Pedersen et al. 2014), which pre-process both reference genome and sequencing reads *in silico* to reduce the number of

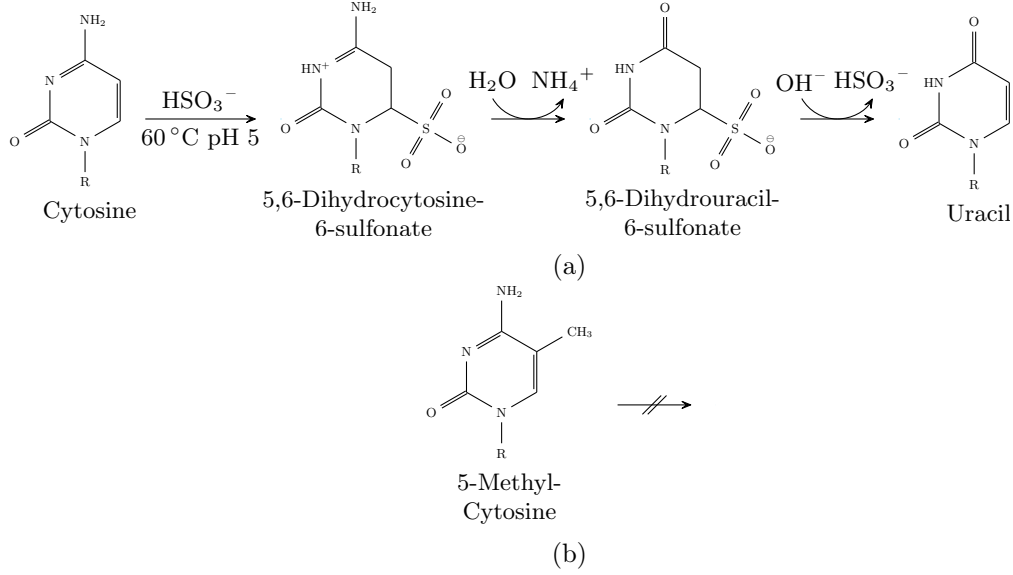


Figure 1.1: Reaction scheme for bisulfite conversion. **(a)** Bisulfite conversion of non-methylated cytosines. i) Acid-catalysed sulfonation, ii) Hydrolytic deamination, iii) Base-catalysed desulfonation; **(b)** Conversion of methylated cytosines is (kinetically) inhibited

possible bases to three depending on strand orientation. The forward strand is converted to contain only thymine, adenine and guanine positions (C→T conversion), whereas on the reverse strand only adenine, thymine and cytosine (G→A conversion) are retained. Out of all four possible alignments between either combination of read and reference strands, the best unique alignment is chosen. The methylation state for each cytosine position in the alignment can then be retrieved by identifying mismatches against the non-converted reference genome.

Assessment of DNA Methylation at Single-Base Resolution

The interplay between methylation and de-methylation of DNA in the nucleus by virtue of antagonistically acting enzymes creates a steady state with a stochastic component. DNA methylation is dynamic in nature across developmental stages (Feng et al. 2010), tissue types (Widman et al. 2014), cell types (Hossain et al. 2017) and is tightly integrated with cell cycle (Borges et al. 2021; Ning et al. 2020). Most of these factors are not controlled for in a typical bisulfite-sequencing experiment. Unless single-cell sequencing is conducted or synchronised cell lines are used, sequencing libraries contain DNA from a heterogenous mixture of cell or even tissue types, in different stages of the cell-cycle at the time of sampling. For these reasons, as well as allele-specific epigenetic phenomena such as imprinting (Köhler and Weinhofer-Molisch 2010) and paramutation (Chandler 2007), cytosine methylation as measured by WGBS is non-binary at a given genomic cytosine position and requires quantitative assessment. Typically, this is achieved by creating pile-up information from the alignment consensus, thereby extracting counts of reads that support methylation (read as C) at each cytosine position and of those that do not (read as T). Given the sum of all reads covering a given position, methylation proportions can

be calculated as $\frac{C}{C+T}$ for each genomic cytosine position and are usually expressed as percentages.

Challenging Aspects in the Identification of Differential Methylation

Data obtained from WGBS experiments imposes some caveats to the downstream analysis and requires careful study design. Insufficient replication and low sequencing depth cause lack of statistical power and can introduce bias due to sequencing error rates. Uneven coverage distribution along the genome, especially at low sequencing depth, can lead to poor estimates of methylation rates because of the resulting increase in variance. Furthermore, data quality can be negatively affected by incomplete conversion through sodium bisulfite treatment of DNA, requiring diligent protocol optimisation to maximise conversion rates on one hand, while avoiding overfragmentation of DNA induced by the harsh chemical conditions during bisulfite conversion on the other. In addition, protocols for library preparations can introduce methylation bias at the read ends if adapter ligation is performed after bisulfite conversion, and often include PCR amplification that can result in overrepresentation of reads originating from genomic regions with certain sequence features (Olova et al. 2018). Common practices to avoid and mitigate these pitfalls during downstream analysis include biological and technical replication, data pre-processing in the form of read trimming as well as removal of PCR duplicates. For plant material, unmethylated chloroplast DNA can be conveniently used as an internal control to assess bisulfite conversion rates, but for other organisms adding appropriate spike-in DNA to the sequencing library is required.

Current Methods for Identification of Differential Methylation

Given a genomic cytosine position, let n denote the total number of reads covering that position, k the number of reads that show methylation at that position and p the probability to observe true methylation at that position. Under the assumption that k follows a binomial distribution (Equation 1.1), p can be estimated from the observed read count proportion $\frac{C}{C+T}$ in a typical WGBS experiment, given sufficiently high coverage.

$$Pr(k|p, n) = \binom{n}{k} p^k (1 - p)^{n-k} \quad (1.1)$$

where:

$$\binom{n}{k} = \frac{n!}{k!(n - k)!} \quad (1.2)$$

While technical replication can mitigate low coverage scenarios to some extent, a binomial model falls short of incorporating other sources of variance typically present in biological datasets such as batch effects, inconsistent bisulfite conversion across samples, and epigenetic variability between samples. To overcome this limitation, current methods for detection of differential methylation (Molaro et al. 2011; Sun et al. 2014; Park and Wu 2016; Feng et al. 2014) approximate methylation status by using a generalised form of the binomial distribution, thereby assuming that the methylation proportion p at position i is a random variable that is drawn from a beta distribution with shape parameters α and β (Equation 1.4).

Consequently, the resulting compound beta binomial distribution (Equation 1.3) allows for joint modeling of both within-sample variance, due to read sampling at given coverage, and between-sample variance, due to biological variability and other confounding factors.

$$Pr(C_i|\alpha, \beta, C_i + T_i) = \binom{C_i + T_i}{C_i} \frac{B(\alpha + C_i, \beta + T_i)}{B(\alpha, \beta)} \quad (1.3)$$

where:

$$B(\alpha, \beta) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{\int_0^1 p^{\alpha-1}(1-p)^{\beta-1} dp} \quad (1.4)$$

While the biological relevance of differential methylation at single loci is uncertain, known epialleles and numerous studies about the influence of promoter methylation on gene expression suggest that changes in DNA methylation are indeed relevant at region-level. To that end, existing methods use different approaches: they either consolidate differentially methylated positions found to be in spatial proximity into regions, pre-select regions based on *a priori* knowledge such as genomic annotation, or segment the genome before assessing differential methylation. However, most of them were developed for studies in mammalian systems and the underlying assumptions do not necessarily hold for the specifics of DNA methylation in plants as will be further outlined in Chapter 2. Critically, significance testing based on beta binomial modeling for millions of genomic cytosine positions is not only computationally demanding, but also leads to a massive multiple testing burden. In the case of some species, *A. thaliana* among them, only a small fraction of the genome space is methylated and when correcting for multiple tests, a severe penalty in detection sensitivity is imposed because a large fraction of tested positions is non-informative. Constraining significance testing to the methylated space of the genome prior to statistical assessment therefore seems favourable. A previously described, unsupervised approach achieves this by implementing a two-state hidden markov model that learns sample-specific parameters of methylation distributions to probabilistically classify and segment methylated regions (Molaro et al. 2011). Chapter 2 will present an adaptation of this very approach to accomodate the complex genetic architecture underlying context-specific DNA methylation in plants as it was outlined in section 1.1.3, and combines the methodology with a testing strategy that enables detection of differential methylation in large sample populations.

1.2 High-Throughput Plant Phenotyping

Assessment of phenotypic traits constitutes an integral component of plant biology, not only because of the emerging role of epigenetic contributions to phenotypic appearance and diversity (Zhang et al. 2013; Xu et al. 2016). Scoring phenotypic outcomes of known mutations, random mutagenesis or allelic variants in natural populations is equally important to plant scientists as the identification of individuals with increased yield, resilience or other favourable traits is to plant breeders. Phenotypic measurements of interest can be manifold, ranging from biomass accumulation, seed production as a proxy for fitness, germination and flowering time to photosynthetic activity or water use efficiency. Some of these traits either require manual intervention to varying extent or specialised

equipment, leading to limited scalability, as is the case for dry or fresh weight measurements, seed weight determination, or scoring of flowering time phenotypes, for example. With increasing demand for high-throughput measurements, image-based plant phenotyping is often employed, in large part because it can be readily automated and also because it is usually applicable across laboratories, greenhouses, or growth-chambers. In field settings, drone-based approaches are often applied (Sankaran et al. 2015). These enable the monitoring of phenotypic traits across many hectares of land but sacrifice on resolution and image quality that is required for fine-grained assessment. Depending on the scale of an experiment, the number of images can rapidly grow beyond a single researchers' capacity to analyse them manually, creating the necessity for automation during the process while at the same time retaining high accuracy. Software tools for image analysis have been developed, some of them offering a graphical user interface (GUI) for assisted, semi-automated evaluation of traits of interest (Schindelin et al. 2012), whereas other solutions aim to keep required user interaction at a minimum (Grady 2006; Rother et al. 2004). A specialised framework for plant phenomics that is actively maintained and developed is PlantCV (Fahlgren et al. 2015; Gehan et al. 2017). It makes extensive use of the OpenCV library (Bradski 2000) and provides its functionality in the form of building blocks for implementation of automated workflows for phenotyping purposes. Typically, the most challenging step towards the extraction of phenotypic measurements from image data is the separation of plant from background in a process called segmentation. The field of Computer Vision has developed a broad variety of solutions to this problem that either employ thresholding (Otsu 1979; Niblack 1985; Sauvola and Pietikäinen 2000), clustering approaches (Lloyd 1982), algorithms based on morphology (Márquez-Neila et al. 2014), contours (Chan and Vese 1999; Caselles et al. 1997), or watershed transformations (Beucher and Lantuéjoul 1979). While these approaches work well for single images or very homogenous datasets, many factors can negatively influence performance and accuracy of segmentation. These factors include, but are not limited to, light exposure, color composition, background configuration, and contrast.

1.2.1 Methods in Plant Phenotyping

Current methods in image-based plant phenotyping often rely on color information to maximise contrast between plant and background. This typically happens through color space conversions, aiming to maximise the distance between green leaves and background in one or more color channels. In practice, based on an empirical threshold value, the converted image is then subjected to binary thresholding, with the aim to obtain a segmentation mask that correctly delineates the plant and at the same time excludes the background. While these approaches are fast and computationally efficient, they require careful, often empirical fine-tuning of parameters to avoid over- or undersegmentation, especially if the image data is not homogenous and has been sourced from different camera setups, resulting in varying light exposure, resolution, or background configuration.

1.2.2 Deep Learning in Plant Phenotyping

Deep learning is a form of supervised machine learning that has gained momentum with the rise of *Big Data*. Initially, it found its main application in data-intensive problems in the realms of text and language processing (Devlin et al. 2018), image (Krizhevsky et al.

2012) and speech recognition (Hinton et al. 2012a). This has changed in recent years. The application of deep neural networks has broadened, and in many disciplines within the natural sciences, including biology and medicine, where the amount of data has grown in abundance, researchers have started to explore the applicability of deep learning to address complex classification and regression problems in their fields. Among those are medical imaging (Bar et al. 2015; Gulshan et al. 2016), drug discovery (Ma et al. 2015; Stokes et al. 2020) and toxicity prediction (Mayr et al. 2016), structural biology (Jumper et al. 2021; Yang et al. 2020), metabolomics (Kantz et al. 2019), long-read sequencing (Wick et al. 2019), and genomics (Poplin et al. 2018; Kelley et al. 2016; Washburn et al. 2019; Lin et al. 2022).

Two main network architectures exist in deep learning, on the one hand there are recurrent neural networks (RNNs) that have turned out to be most useful for sequential data. On the other hand there are feed-forward neural networks (FNNs), including convolutional neural networks (CNNs) that exhibit state-of-the-art performance for classification tasks in image, audio, and speech recognition. The remainder of this section shall focus on the latter architecture, give some historical remarks about how FNNs came to be and allude to the usefulness of CNNs for image-based plant phenotyping.

An Historical Overview of Feed-Forward Neural Networks

Historically, the origins of deep learning date back to the first mathematical model describing an artificial neuron (McCulloch and Pitts 1943), inspiring the conceptualisation of the perceptron fifteen years later (Rosenblatt 1958). The concept of the perceptron still represents the basic building block of modern artificial neural networks and can be formalised as a linear threshold unit (Equation 1.5).

$$f(x_i; W, \theta) = \begin{cases} 1 & \text{if } W^T x_i \geq \theta \\ 0 & \text{if } W^T x_i < \theta \end{cases} \quad (1.5)$$

Several features of this linear classifier simplistically resemble those of a neuron. The inputs x_i can be viewed in analogy to the electrical input generated by sequential membrane de-, re- and hyperpolarisation of connected axons, that ultimately result in an action potential. The output of the activation function (Equation 1.5) then depends on whether or not a threshold θ is exceeded by that signal, closely resembling the biological case, in which a neuron fires or not. In both cases, the final outcome is dependent on the weights W^T , that take the form of matrices and figuratively resemble properties of the synaptic interface, such as the presence or absence of neurotransmitters and their concentrations. The perceptron marked the beginning of artificial neural networks as we know them today, but it took time to overcome its initial shortcomings. Research in that direction was *de facto* halted for almost two decades after seminal work in the field had shown that the use of perceptrons in classification was limited to linearly separable data (Minsky and Papert 1969). While multilayer perceptrons (MLPs) were already considered a solution to this problem by introducing additional intermediate, later called hidden layers. However, connecting multiple perceptrons in such a layered network was deemed computationally infeasible at the time. It was not until the 1980s when, coinciding with the unfolding of Moore’s law, capabilities of central processing units (CPUs) had sufficiently increased and the study of feed-forward neural networks (FNNs) gained renewed interest. An algorithm

for efficient training of artificial neural networks through backpropagation of error had been proposed (Werbos 1982; Rumelhart et al. 1986) and applied to MLPs (Rumelhart et al. 1986). The algorithm is based on the idea to change the activation function from a discontinuous step function that yields a binary outcome (Equation 1.5), to a continuous, differentiable function that is related to the input in any network layer. By that, it became possible to iteratively update layer weights based on the error between labeled training data and the classification estimate of the network, effectively turning the training process into an optimisation problem. A loss function that quantifies this error is iteratively minimised by *gradient descent*, calculating partial derivatives with respect to all network weights using the chain rule. After each iteration, this allows the weight matrices throughout the network layers to be updated in opposite direction to the gradient, enabling network performance to incrementally improve as training progresses. Backpropagation of error paved the way for larger networks with more depth in the sense of multiple hidden layers that gave rise to the term *deep neural networks*. Apart from their computational complexity, increased network depth immediately posed new challenges. Activation functions used early on, such as the logistic function, suffered from the problem that their derivatives take near-zero values in the areas of saturation. As the gradients multiply according to the chain rule, increasing network depth was thus associated with decreasing signal strength. Solutions to address the problem of vanishing gradients include novel activation functions (Nair and Hinton 2010; Clevert et al. 2015), normalization techniques (Ioffe and Szegedy 2015; Ba et al. 2016; Salimans and Kingma 2016), improved weight initialisation (Glorot and Bengio 2010; Saxe et al. 2013), as well as gating in RNNs (Hochreiter and Schmidhuber 1997). Another problem with larger networks stems from the fact that the weight matrices are high-dimensional and the loss function non-convex, meaning that its optimisation converges to local minima only. To avoid the fallacy of overfitting, in which classification works well on training and validation data and specific characteristics thereof but not on *unseen* data, both pre-existing and newly-developed regularisation schemes have been applied to neural networks (Tibshirani 1996; Hoerl and Kennard 1970; Hinton et al. 2012b; Ioffe and Szegedy 2015). Despite these improvements, the requirement for a large and diverse set of training data remains crucial for obtaining a network capable of good generalisation. With the rise of *Big Data*, this requirement can well be satisfied and the applications of deep learning continue to grow in numbers. Neural networks can handle vast amounts of high-dimensional data and can serve a broad set of classification tasks. Technological advances in processing power on one hand, particularly in the design of graphics processing units (GPUs), have enhanced the feasibility of large networks. GPUs are, with their unprecedented parallelism and memory bandwidth, especially suitable for the vector and matrix multiplications that underlie the training process in deep learning, making the process more time-efficient compared to CPU-based training by orders of magnitude. Substantial investments on the other hand, not only in the sense of scholarly research but also resulting from commercial interest in data processing by large technology companies like Google or Facebook, have driven the development of a number of deep learning frameworks, such as *Keras* (Chollet et al. 2015), *TensorFlow* (Abadi et al. 2016), *PyTorch* (Paszke et al. 2019) or *Flux.jl* (Innes et al. 2018) all with extensive documentation, support for multiple programming languages, and active user communities. These represent a non-negligible and certainly accelerating aspect towards the spread and adoption of deep learning, as they greatly facilitate the application of the methodology by lowering the bar of entry for non-experts in the subject matter, biologists among them.

Convolutional Neural Networks (CNNs)

A specialised form of FNNs primarily used in image classification, but also used for video, speech and audio data, are CNNs (LeCun et al. 1989; Lecun et al. 1998). Similar to the case of the perceptron outlined above, the architecture of CNNs can be envisioned in close analogy to neurophysiological processes in the visual cortex. There, individual cortical nerve cells are only responsive to stimuli within a spatially restricted area known as the receptive field (Hubel and Wiesel 1959). The visual field is then constructed from multiple receptive fields that overlap between cells in close proximity. Conceptually, in convolutional layers of CNNs, such receptive fields are realised by means of filters represented by a weight matrix kernel of smaller size relative to the input. The filter weights are shared, meaning they have constant values that are applied to all positions of the input, resulting in a so-called convolution over the input. The output of a convolutional layer is then transformed with a non-linear activation function, resulting in feature map representations that are passed on to the next layers in the network, typically additional convolutional layers, optional pooling or subsampling layers, followed by fully-connected layers (Krizhevsky et al. 2012). The latter receive input from all units in the previous layer, again closely resembling the neurophysiological construction of the visual field. CNNs have proven particularly useful for data that contains local structure, as is the case in digital images for example, where pixels in spatial proximity tend to be highly correlated. Basic patterns such as corners, edges, textures and certain shapes represent recurring structures throughout the image, allowing to successful application of the concept to small receptive fields without sacrificing too much information. This is reflected by the superior role that CNNs take in the yearly ImageNet Large Scale Visual Recognition Challenge (ILSVRC) (Russakovsky et al. 2015) ever since their first application in 2012, when the AlexNet network architecture outperformed competing non-CNN architectures by a large margin (Krizhevsky et al. 2012).

Existing Applications of Deep Learning in Plant Phenotyping

Compared to the classical approaches in plant phenotyping outlined in Section 1.2.1, which often rely on color information alone, deep learning has the potential to overcome the associated shortcomings of those methods. Because the classification output of CNNs incorporates patterns such as shapes, textures, edges, and more complex structures in addition to color information, they are well suited for cases in which the images to be analysed are heterogenous, differ in light exposure, were sourced from different imaging platforms, or have complex background configurations. This has led to a range of applications in plant phenotyping, targeting different tissue types such as leaves (Ubbens et al. 2018; Ward et al. 2018) and roots (Wang et al. 2019; Douarre et al. 2018), reproductive tissues such as pollen (Olsson et al. 2021; Sevilano et al. 2020), fruits (Feldmann et al. 2020; Hamidinekoo et al. 2020) and seeds (Braguy et al. 2021; Warman et al. 2021). Notably, deep learning is also increasingly used in plant pathology to classify disease symptoms and lesions (Stewart et al. 2019; Polder et al. 2019; Singh et al. 2019).

Transfer learning

As mentioned earlier, obtaining a network with good generalisation typically requires a large and diverse dataset to avoid overfitting to image artifacts, background composition,

or certain color patterns. Parts of these potential pitfalls can be mitigated by artificially increasing training data through augmentation of existing images. Color channel normalisation, random cropping, mirroring, or tilting can help to increase variability within the dataset; however, these techniques fail to address another central aspect of network training - that is computational complexity. Incremental weight improvements through minimisation of a loss function, as they happen during the training process, require many iterations until convergence. In order to train large networks within reasonable timeframes, this goes along with substantial hardware requirements, especially with regard to GPUs. The availability of vast amounts of labeled training data provided by publicly available datasets (Deng et al. 2009; Lin et al. 2014; Everingham et al. 2010) not only led to the development of refined network architectures over time (Szegedy et al. 2014; He et al. 2015; Chollet 2016), but also enabled the application of transfer learning. Transfer learning is a strategy that builds upon pre-trained networks, which already contain useful layer weights for feature extraction acquired over many training iterations on existing datasets with often careful parameter optimisation. The widely used ImageNet dataset (Deng et al. 2009) consists of millions of hand-annotated images, representing a comprehensive resource for training and evaluation of deep learning models that serve classification tasks such as object detection or image segmentation. Utilising such resources through transfer learning can greatly reduce the amount of annotations required to retrain a network for specialised classification tasks (Razavian et al. 2014; Donahue et al. 2013), largely increasing feasibility of deep learning methodology for a wide range of applications, including high-throughput plant phenotyping, as will be presented in Chapter 5.

2 MethylScore, a Pipeline for Accurate and Context-Aware Identification of Differentially Methylated Regions from Population-Scale Plant WGBS Data

status	<i>under review - Quantitative Plant Biology</i> bioRxiv preprint (2022) doi:10.1101/2022.01.06.475031
authors	Patrick Hüther* (P.H.), Jörg Hagmann* (J.H.), Adam Nunn (A.J.), Ioanna Kakoulidou (I.K.), Rahul Pisupati (R.P.), David Langenberger (D.L.), Detlef Weigel (D.W.), Frank Johannes (F.J.), Sebastian J. Schultheiss (S.J.S.) and Claude Becker (C.B.) * <i>equal contributions</i>
contributions	P.H., J.H., F.J., D.W., and C.B. conceived the study and developed the pipeline; D.L. and C.B. supervised the work. J.H. wrote the MethylScore code; P.H. built the Nextflow pipeline. I.K. generated the epiRIL data. P.H. and A.N. analysed the data. R.P. prepared methylation calls on the <i>A. thaliana</i> 1001 Genomes dataset. P.H. and C.B. wrote the manuscript.

bioRxiv preprint doi: <https://doi.org/10.1101/2022.01.06.475031>; this version posted January 6, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

MethylScore, a pipeline for accurate and context-aware identification of differentially methylated regions from population-scale plant WGBS data

Patrick Hühner^{1,2,*}, Jörg Hagmann^{3,*}, Adam Nunn^{4,5}, Ioanna Kakoulidou⁶, Rahul Pisupati¹, David Langenberger⁴, Detlef Weigel⁷, Frank Johannes^{6,8}, Sebastian J. Schultheiss^{3,§}, & Claude Becker^{1,2,§}

¹ Gregor Mendel Institute of Molecular Plant Biology GmbH, Austrian Academy of Sciences, Vienna BioCenter (VBC), 1030 Vienna, Austria

² LMU Biocenter, Faculty of Biology, Ludwig-Maximilians-University Munich, 82152 Martinsried, Germany

³ Computomics GmbH, 72072 Tübingen, Germany

⁴ ecSeq Bioinformatics GmbH, 04103 Leipzig, Germany

⁵ Department of Computer Science, Leipzig University, 04107 Leipzig, Germany

⁶ Department of Plant Sciences, Technical University of Munich, Freising, Germany

⁷ Department of Molecular Biology, Max Planck Institute of Developmental Biology, 72076 Tübingen, Germany

⁸ Institute for Advanced Study, Technical University of Munich, Garching, Germany

* These authors contributed equally to this work.

§ To whom correspondence should be addressed: sebastian.schultheiss@computomics.com; claud.becker@biologie.uni-muenchen.de

Abstract

Whole-genome bisulfite sequencing (WGBS) is the standard method for profiling DNA methylation at single-nucleotide resolution. Many WGBS-based studies aim to identify biologically relevant loci that display differential methylation between genotypes, treatment groups, tissues, or developmental stages. Over the years, different tools have been developed to extract differentially methylated regions (DMRs) from whole-genome data. Often, such tools are built upon assumptions from mammalian data and do not consider the substantially more complex and variable nature of plant DNA methylation. Here, we present MethylScore, a pipeline to analyze WGBS data and to account for plant-specific DNA methylation properties. MethylScore processes data from genomic alignments to DMR output and is designed to be usable by novice and expert users alike. It uses an unsupervised machine learning approach to segment the genome by classification into states of high and low methylation, substantially reducing the number of necessary statistical tests while increasing the signal-to-noise ratio and the statistical power. We show how MethylScore can identify DMRs from hundreds of samples and how its data-driven approach can stratify associated samples without prior information. We identify DMRs in the *A. thaliana* 1001 Genomes dataset to unveil known and unknown genotype-epigenotype associations. MethylScore is an accessible pipeline for plant WGBS data, with unprecedented features for DMR calling in small- and large-scale datasets; it is built as a Nextflow pipeline and its source code is available at <https://github.com/Computomics/MethylScore>.

Introduction

Cytosine methylation, which is often used synonymously with DNA methylation, describes the covalent attachment of a methyl group to carbon 5 of cytosine, resulting in 5-methylcytosine (5mC). In plants, as in most eukaryotes, DNA methylation is part of an epigenetic mechanism involved in transposon silencing (Miura et al., 2001), heterochromatin formation (Lippman et al., 2004), and gene regulation (Jaenisch and Bird, 2003). It also plays a role in genome organization (Zemach et al., 2013; Huff and Zilberman, 2014), regulation of development (Finnegan et al., 1996; Papareddy et al., 2021; Ronemus et al., 1996) and imprinting (Jullien et al., 2008; Pignatta et al., 2018; Gehring et al., 2006). DNA methylation can also be a source of phenotypic variation in natural populations (Schmitz et al., 2013; Eichten et al., 2011; Heyn et al., 2013).

Despite recent advances in long-read sequencing, short-read WGBS is still considered the gold standard for analysis of DNA methylation at single-base resolution. Treatment of genomic DNA with sodium bisulfite causes hydrolytic deamination of unmethylated cytosine (C) into uracil, which upon sequencing library amplification by PCR is converted to thymine (T) (Frommer et al., 1992). The initial step is kinetically inhibited in the case of 5mC, and so methylated cytosines will be interpreted normally during short-read sequencing, whereas unmethylated, converted cytosines will be interpreted as T by the base caller. Comparison to the known reference genome can subsequently reveal the methylation state of each cytosine.

Features of plant DNA methylation

In many eukaryotic organisms, cytosines can be in a methylated state when they are in a CG sequence context, i.e., followed by a guanine (G). In mammalian genomes, for example, the majority of CGs are methylated, while unmethylated cytosines are grouped in so-called CpG islands, which are non-randomly distributed along the genome and play an important role in transcriptional regulation (Deaton and Bird, 2011). In plant genomes, however, the situation is fundamentally different and can vary substantially from one species to another. DNA methylation in plants occurs in three possible sequence contexts: CG, CHG, and CHH (where H is any base but G) (Law and Jacobsen, 2010). Each type of methylation is established, maintained and regulated by a specific molecular machinery (Stroud et al., 2013b; Henderson and Jacobsen, 2007; Law and Jacobsen, 2010). The frequency of methylation along the genome differs by sequence context; for example, in the model plant *Arabidopsis thaliana*, CGs are methylated most frequently relative to the total number of cytosines in that context (~24% of all CGs are 5mCGs), followed by CHG (~7%) and CHH (~1.7%) (Cokus et al., 2008).

While DNA methylation in CG and CHG is symmetric, i.e., the complementary sequences on the opposite strand mirror the methylation state, CHH methylation is asymmetric and occurs on just one strand at that particular locus. For CG methylation, the information retained by semi-conservative DNA replication facilitates re-establishment of methylation on the daughter strand via the methyltransferase METHYLTRANSFERASE 1 (MET1) that, akin to its mammalian homologue DNA METHYLTRANSFERASE 1 (DNMT1), shows substrate preference to hemimethylated DNA (Kankel et al., 2003; Jeltsch, 2006; Saze et al., 2003). In contrast, maintenance of CHG and asymmetric CHH methylation depends on additional factors such as histone modifications and small RNAs. Methylation of CHG cytosines by CHROMOMETHYLASE 3 (CMT3) reciprocally depends on H3K9 dimethylation (Jackson et al., 2002; Du et al., 2012), while CHH methylation requires CHROMOMETHYLASE 2 (CMT2)

bioRxiv preprint doi: <https://doi.org/10.1101/2022.01.06.475031>; this version posted January 6, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

(Stroud et al., 2014; Zemach et al., 2013) and DOMAINS REARRANGED METHYLTRANSFERASE2 (DRM2) recruited through small-RNA-dependent RNA-directed DNA methylation (RdDM) (Aufsatz et al., 2002; Cao et al., 2003; Cao and Jacobsen, 2002); (Wassenegger et al., 1994; Matzke and Mosher, 2014).

It is important to note that — at least in some plant species, including *A. thaliana* — DNA methylation in the different sequence contexts also differs in terms of methylation rate, i.e., in the consistency of the methylation status of a given cytosine across different cells. CG cytosines tend to have a binary methylation state, being either always unmethylated or almost always methylated in different cells of a given tissue. In contrast, cytosines in CHG and CHH show more variable methylation status, with mean methylation rates across all methylated CHG and CHH cytosines of ~50% and ~30%, respectively. This has consequences for the analysis of differential methylation between samples (Becker et al., 2011), for example in software which would model the underlying methylation status based on expected distributions (Hansen et al., 2012; Hebestreit et al., 2013; Korthauer et al., 2018).

Whole-genome DNA methylation analysis as a means to understand natural variation and stress response

Since the first characterizations of the whole-genome DNA methylation profile in *A. thaliana* (Cokus et al., 2008; Lister et al., 2008), there has been a growing interest in studying this epigenetic mark at the genomic level to better understand developmental processes (Pignatta et al., 2018; Manning et al., 2006), stress responses (Wibowo et al., 2016; Liu et al., 2018), phenotypic plasticity, and natural variation (Schmitz et al., 2013; Kawakatsu et al., 2016). For example, heritable genetic variation alone is not always sufficient to explain the range of phenotypic diversity observed for individuals of the same species, and epigenetic variation is likely to account for at least some of the missing heritability (Manolio et al., 2009). Epigenetic variation, albeit often confounded by genetic variation, has been recognized as a source of natural diversity (Riddle and Richards, 2002; Cervera et al., 2002; Vaughn et al., 2007). As a consequence, some phenotypic traits, including floral symmetry (Cubas et al., 1999), fruit development and morphology (Manning et al., 2006; Ong-Abdullah et al., 2015; Zhong et al., 2013), plant height (Miura et al., 2009), and resistance to pathogens (Liégard et al., 2019), have been associated with naturally occurring epigenetic alleles (epialleles) that do not appear to be due to linked genetic variation. In addition, studying the conditional dynamics of DNA methylation at the whole-genome level has highlighted the contribution of epigenetic regulation to stress response and tolerance (Wibowo et al., 2016).

The challenges in determining differential DNA methylation

DNA methylation is highly dynamic, and there is substantial variation among individuals or (sub)populations. It is important to distinguish specific, relevant differences from stochastically occurring methylation differences between biological replicates. On a genome-wide scale, this is challenging without *a priori* knowledge of stratification, either based on sequence context or by sample grouping. Most experimental studies contrast DNA methylation from different samples to each other, be it mutant background and wild type, treatment and control, or different natural accessions. Statistical comparisons between samples aim to identify DNA methylation differences at either the single-cytosine (differentially methylated positions; DMPs) or the region (differentially methylated regions; DMRs) level. While DMPs provide useful information on the rate at which epigenetic changes occur (Becker et al., 2011; Schmitz et al.,

2011; van der Graaf et al., 2015), DMRs are arguably more relevant in a functional biological context because they can affect contiguous stretches of DNA and hence potentially influence the accessibility of regulatory elements. However, the nature of WGBS data imposes several caveats to accurately determining DMPs and DMRs. Some of these reside in the experimental design or the quality of the sequencing library. Insufficient replication, for example, limits the statistical power of differential analyses. Uneven coverage can bias the base calling because of sequencing error rates, while incomplete bisulfite conversion can cause a false estimate of methylated cytosines, as can duplicated reads that arise from library over-amplification.

The statistical analysis of differential methylation often brings along another set of caveats. For example, strategies that are based on defining DMRs as clusters of spatially adjacent DMPs, such as *DSS* (Feng et al., 2014; Ziller et al., 2013; Schultz et al., 2015; Feng et al., 2014), are subject to a heavy multiple testing burden resulting from the large number of cytosines in the genome that need to be tested individually. The same applies to window- or sliding-window-based approaches, as implemented, for example, in *methylKit* (Akalin et al., 2012). In contrast, strategies that call DMRs only in pre-defined regions, e.g., in annotated features, risk missing relevant loci in the analysis.

More recently developed tools for DMR calling are centered around pre-selection of genomic regions. *metilene* (Jühling et al., 2016), *dmrseq* (Korthauer et al., 2018) and *HOME* (Srivastava et al., 2019) implement multi-step strategies that first restrict the testable genome space to candidate regions with evidence of methylation, prior to assessing significant differences. Even though *HOME* takes non-CG methylation into account, all three of these tools have been developed with mammalian (mostly human) DNA methylation data in mind and are therefore built on assumptions, including most cytosines being methylated, strong local correlation between methylation states, predominant (or exclusive) CG methylation, and largely binary methylation states, that do not reflect the distinct genetic control underlying the CG, CHG and CHH methylation contexts in plants. This implies that if these assumptions are indiscriminately applied, many methylated cytosines and regions are likely to be falsely discarded.

Here, we set out to address this gap by developing a tool for the robust identification of DMRs from plant WGBS data that takes into account the complexity and variability of plant DNA methylation, while using an informed, restricted set of candidate regions for statistical testing. Furthermore, we aimed to avoid the necessity for pre-defining sample groups, required by existing tools, to increase the applicability to sample populations with inherent group structure and to prevent experimenter bias. We present MethylScore, a convenient and stable pipeline that enables biology researchers with little computational background to process WGBS data, from read alignments to DMR output. The differential methylation analysis module of MethylScore is built around the two-state Hidden-Markov-Model-based approach described by (Molaro et al., 2011). To identify and segment methylated regions in plant genomes, independent of prior information, we extended the original implementation beyond the CG sequence context, allowing the algorithm to train distinct parameter sets for each methylation context.

MethylScore trains on the different properties of CG, CHG, and CHH methylation by estimating parameters of a beta-binomial distribution for each sequence context, accounting for both stochastic variance in the coverage distribution (assumed to be beta distributed) as well as between-sample biological variance (binomially distributed). This means that MethylScore learns from the actual WGBS data, thus avoiding potentially erroneous or species-specific

bioRxiv preprint doi: <https://doi.org/10.1101/2022.01.06.475031>; this version posted January 6, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

assumptions about the distribution of methylation or arbitrary thresholds. This way, the number of statistical tests is constrained to regions of interest, curtailing the multiple-testing problem. Moreover, its built-in population-scale approach allows identifying DMRs in large datasets, without prior information on sample groups, enabling an unbiased DMR calling. Using publicly available datasets from *A. thaliana* (1001 Genomes Consortium, 2016; Kawakatsu et al., 2016; Tedeschi et al., 2019; Ning et al., 2020; Zhang et al., 2021; 1001 Genomes Consortium, 2016; Kawakatsu et al., 2016; Wibowo et al., 2018) and rice (Stroud et al., 2013a), we show that MethylScore is able to segment plant genomes with very different global DNA methylation profiles. In absence of sample information, MethylScore identified group-specific DMRs and was able to detect population signals in datasets with hundreds of samples. Ultimately, we used the DMRs thus identified in the *A. thaliana* 1001 Genomes and Epigenomes datasets (1001 Genomes Consortium, 2016; Kawakatsu et al., 2016) to detect known and unknown genotype-epigenotype associations. MethylScore is built as a Nextflow pipeline, allowing ease-of-access and broad usability; the pipeline can be downloaded from <https://github.com/Computomics/MethylScore>.

Results

Outline of the modular MethylScore pipeline to call DMRs from plant WGBS data

MethylScore is implemented as a modular pipeline that is built in Nextflow (Di Tommaso et al., 2017) for maximal ease-of-access (Figure 1; Materials and methods). It uses reference-aligned sequences in *bam* format or tabular single-cytosine information in *bedGraph* format as they are produced by common WGBS primary analysis tools such as *bismark* (Krueger and Andrews, 2011) or *MethylDackel* (<https://github.com/dpryan79/MethylDackel>). The MethylScore pipeline performs the following general steps: (i) calling of methylated cytosines and determining the per-cytosine methylation rate, (ii) identifying contiguous regions with high DNA methylation in each sample, (iii) determining segments that for statistical testing, and (iv) performing statistical tests to identify DMRs. To start the pipeline, users need to provide a two-column sample sheet with sample identifiers and file paths to the respective *bam* or *bedGraph* files as well as to the reference genome sequence in *fasta* format. If multiple lines in the sample sheet share the same sample identifier, MethylScore will assume technical replication and merge the files prior to further processing.

First, the *samtools sort* (Li et al., 2009) function sorts the alignments by genomic coordinates. As many library preparation protocols for bisulfite sequencing incorporate a PCR amplification step during the protocol, alignments can be optionally processed with *picard MarkDuplicates* (Picard toolkit, 2019) to avoid potential bias by non-uniform read amplification. For the next steps up to the methylated region calling, the alignments are split by chromosome or contig, facilitating full parallelisation of downstream processing, which is open to user configuration in order to suit the available computational resources. *MethylDackel extract* (<https://github.com/dpryan79/MethylDackel>) is used for tabulation of single-cytosine information per sample.

The consensus files for each sample and chromosome are merged into a uniform file format containing methylation information for all cytosines across all samples, together with information about sequence context and strand information for each cytosine. After merging individual chromosomes, the resulting genome matrix file is compressed with *bgzip* and

indexed with *tabix*, allowing for targeted queries of regions from the compressed file. The genome matrix serves as the input for iterative training of one Hidden Markov Model (HMM) per sample. It is implemented as a two-state (unmethylated/methylated) model, designed to learn three context-specific methylation rate distributions by fitting distinct beta-binomial distributions, thus reflecting the underlying biologically distinct control of these sequence contexts in plants (see Materials and methods).

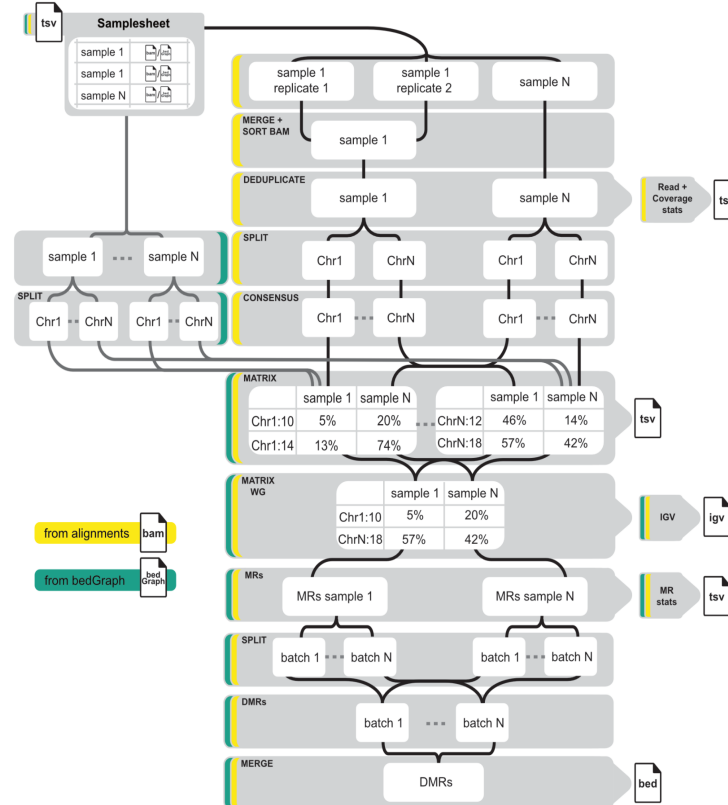


Figure 1. Schematic overview of the MethylScore pipeline.

Schematic overview of processing steps (grey). Different workflow entry points based on the type of input data are indicated as coloured borders. The workflow can be started from genomic alignments (*BAM* format; yellow) or tabular single-cytosine methylation information (*bedGraph* format; green). Chr: chromosome; DMR: differentially methylated region; MR: methylated region.

Based on the resulting parameterization of the trained model, the genome space is segmented into methylated and unmethylated regions by probabilistic classification, yielding genomic coordinates of MRs in *bed* format. These regions are split into batches to facilitate the parallel processing of the ensuing statistical testing strategy. Candidate DMRs in MR batches are preselected based on changes in MR frequency across the sample population for the detection of plausible region boundaries (Supplementary Figure 1). Using an iterative *k*-means algorithm to determine the most suitable number of groups (MacQueen, 1967), samples are clustered according to their mean methylation rates within candidate DMRs. Clusters for each sequence context are assessed for statistical significance by employing log-likelihood ratios based on

bioRxiv preprint doi: <https://doi.org/10.1101/2022.01.06.475031>; this version posted January 6, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

observed read counts under the beta-binomial assumption and testing against a chi-squared distribution with six degrees of freedom as outlined in (Hagmann et al., 2015), followed by FDR control using Storey's method (Storey and Tibshirani, 2003). In a last step, DMRs that pass the FDR-corrected significance threshold are merged, yielding one *bed* file per methylation context containing DMRs between sample clusters.

Segmentation of plant genomes with different DNA methylation composition

We first wanted to explore how MethylScore would segment plant genomes with very different global DNA methylation configurations. In *A. thaliana*, DNA methylation is unevenly distributed, with most methylated cytosines located in the centromeric and pericentromeric regions (Niederhuth et al., 2016). Overall, 11% of cytosines are methylated, with 31.7% of CG, 16.3% of CHG, and 6.2% of CHH cytosines showing methylation rates above zero. In contrast, in the genome of rice (*Oryza sativa*), DNA methylation is more evenly distributed along the chromosomes and more frequent in general, showing 48% CG, 25.3 % CHG and 5.7% CHH methylation. When segmenting WGBS data for both species, MethylScore detected 42,478 methylated regions (MRs) in *A. thaliana* vs. 379,227 in rice, covering 19.5% and 32.5% of the total genome sequence, respectively. Accordingly, the median length of MRs in Arabidopsis (111 bp) was much shorter than in rice (188 bp) (Supplementary Figure 2).

MethylScore identifies differentially methylated regions without a priori sample information

We designed MethylScore to be unaware of sample relationships, which means that it does not require any *a priori* information about replicate groups and that the analysis will not be biased by potentially inaccurate assumptions about sample similarities. Instead, for each candidate DMR, MethylScore groups all samples using an iterative *k*-means approach (see Materials and methods), followed by a beta-binomial-based test to statistically assess whether the group means are significantly different from each other. As a result, grouping may not necessarily reflect the preconceived biological design of the experiment for each candidate DMR, yet local grouping of DMRs has the advantage that it allows to identify patterns in the underlying sample structure that might be overlooked when defining groups beforehand. Averaged over all candidate DMRs, methylation differences should be apparent between, for example, treatment groups or developmental stages *if* the underlying hypothesis for these groups to differ in DNA methylation was correct. To test if MethylScore accurately differentiates between-group and within-group differences in this unsupervised approach, we applied it to *A. thaliana* datasets that had previously been described as having substantial levels of between-group methylation divergence.

We first analyzed a relatively simple WGBS dataset of Columbia-0 (Col-0) wild type and two loss-of-function mutants of *EFFECTOR OF TRANSCRIPTION 1* (*et1-1*) and *EFFECTOR OF TRANSCRIPTION 2* (*et2-3*), respectively, as well as of an *et1-1 et2-3* double mutant, with three biological replicates per genotype. These mutants had previously been shown to have local DNA methylation changes compared to wild-type but also compared to each other (Tedeschi et al., 2019). When running MethylScore on all samples without group replicate information, methylation differences in DMRs accurately reflected the genotypic relationship, indicating that the grouping of samples over all DMRs was driven by methylation differences between genotypic groups (Figure 2). In accordance with the original publication, grouping was most conspicuous in CG and CHG contexts (Figure 2).

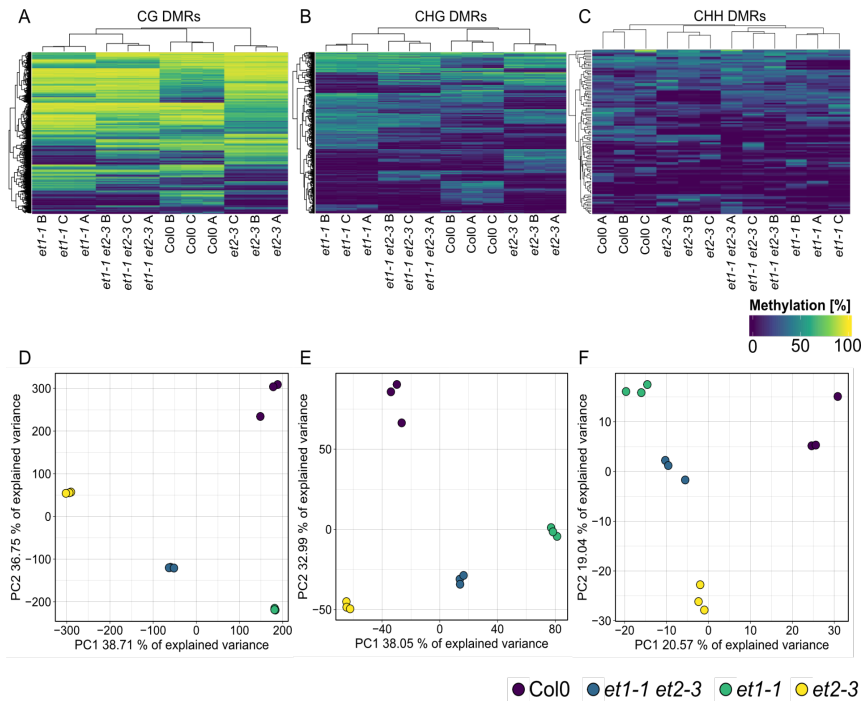


Figure 2: MethylScore DMR calling clusters samples according to genotype.

Heatmaps and Principal Component Analyses (PCAs) of mean methylation rates in 9,487 CG- (A,D), 1,282 CHG- (B,E) and 741 CHH- (C,F) context-specific DMRs identified by MethylScore from WGBS data from flowers of *A. thaliana* Col-0 wild-type, respective *et1-1* and *et2-3* single mutants, and *et1-1 et2-3* double mutants. Original data from Tedeschi et al. (2019) (ENA accession PRJEB12413).

Next, we asked whether MethylScore would accurately classify methylation pathway mutants with more pronounced DNA methylation differences to wild-type. Some of these mutants have substantial global loss of cytosine methylation; we therefore wanted to see whether the training of the HMM, which takes into consideration the methylation rate distributions of each sample, might be affected by this. We used a dataset comprising loss-of-functions mutants of the CHG-specific DNA methyltransferase CMT3, the chromatin remodeler DECREASE IN DNA METHYLATION 1 (DDM1), two regulators of these genes, TESMIN/TSO1-LIKE CXC DOMAIN-CONTAINING PROTEIN 5 (TCX5) and TCX6, and of combinations thereof (Ning et al., 2020). Similar to the *et* mutant analysis (Figure 2), MethylScore accurately clustered the samples according to genotype based on DNA methylation rates within DMRs (Figure 3). Context-specific DMRs also resolved CG- and CHG-specific loss of methylation in *ddm1* and *cmt3*, respectively (Figure 3), indicating that deviations from the standard methylation rate distributions did not affect MethylScore analysis.

bioRxiv preprint doi: <https://doi.org/10.1101/2022.01.06.475031>; this version posted January 6, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

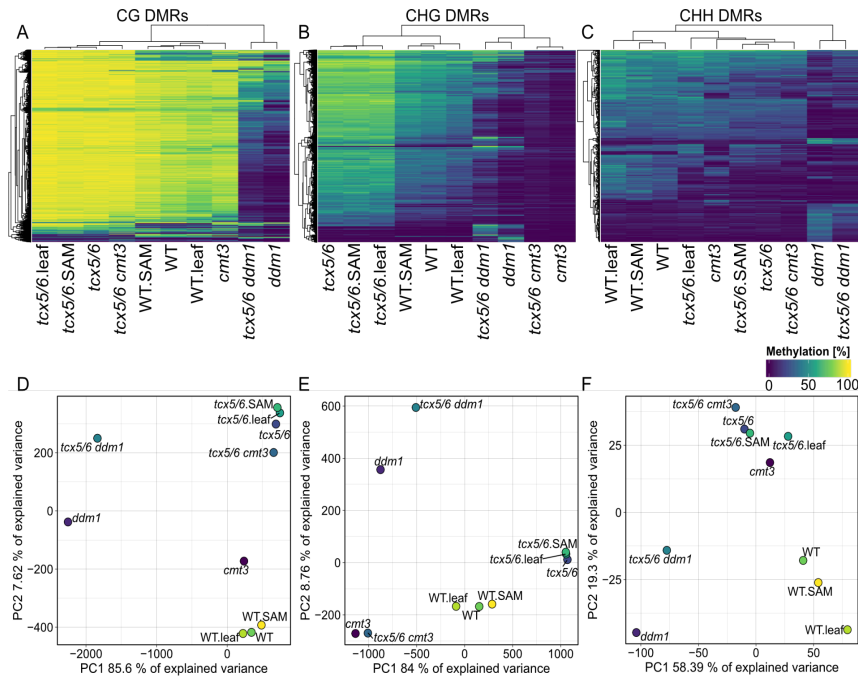


Figure 3: Unsupervised DMR calling from WGBS data of DNA methylation pathway mutants. Heatmaps and PCAs of mean methylation rates in 59,153 CG- (A,D), 63,385 CHG- (B,E) and 440 CHH- (C,F) context-specific DMRs identified by MethylScore from WGBS sequencing data of DNA methylation pathway mutants. The dataset included *ddm1* and *cmt3* single mutants, *tcx5/6* double mutants, as well as *tcx5/6 ddm1* and *tcx5/6 cmt3* triple mutants. DNA had been sampled from leaves and shoot apical meristem (SAM). Original data from Ning et al. (2020); GEO accession GSE137754.

Clustering of methylated regions across regenerated A. thaliana populations underlines partial maintenance of organ-specific methylation profiles in CG

Mutants known to be affected in DNA methylation arguably are expected to present relatively simple cases for genome-wide differential analyses. To test more complex situations, we applied MethylScore's population-scale approach to more complex datasets with less predictable group structure. We re-analysed a published dataset of *A. thaliana* regenerants derived from different tissues of origin that had been obtained by induction of somatic embryogenesis (Wibowo et al., 2018). Instead of employing a pairwise testing strategy between condition groups as it had been pursued in the original study, we used MethylScore to preselect candidate regions based on MR frequency changes across the sample population, and hypothesized that such a strategy should naturally find clusters of regenerants with a similar epigenetic setup, thus allowing to differentiate between organ-derived methylation signatures. The data contained three potentially interacting factors that could contribute to epigenetic variation: regeneration via somatic embryogenesis vs. sexual reproduction; tissue of origin of the somatic embryos (RO: root origin; LO: leaf origin); and tissue of DNA sampling (leaf or root) in the progeny of the regenerants.

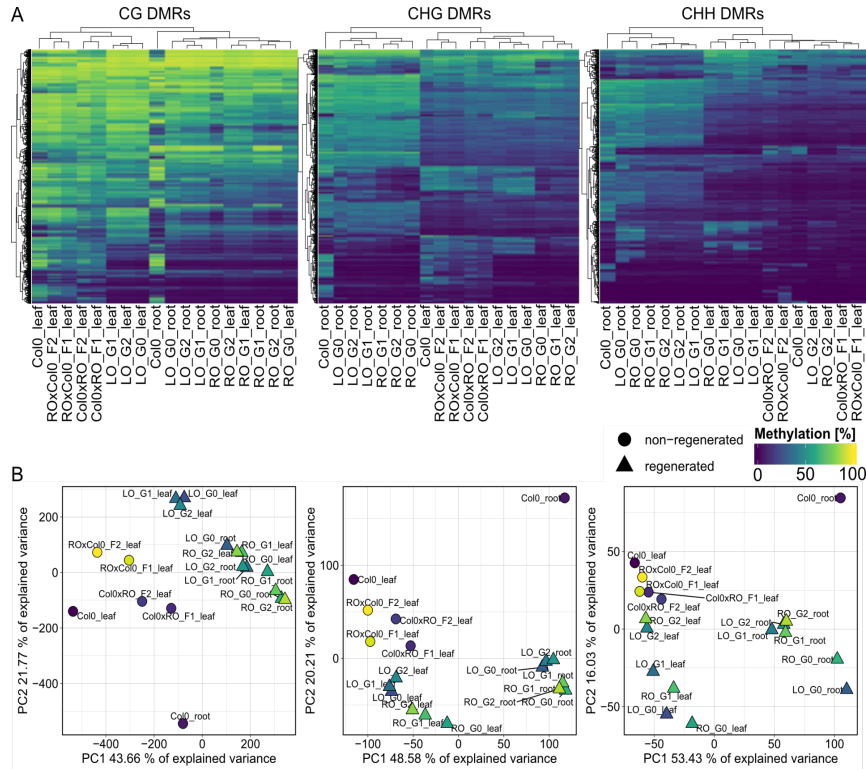


Figure 4: MethylScore population clustering partially reflects epigenetic origin of regenerated plant lineages.

A) Heatmaps show methylation rate averages in regions identified as differentially methylated in CG, CHG, or CHH methylation contexts (from left to right). The dataset includes leaf and root tissue from Col-0 control plants as well as from generation 1 (G1) and generation 2 (G2) progeny of somatic regenerants from root origin (RO) and leaf origin (LO) somatic embryos, and leaf tissue from F1 and F2 backcrosses of RO and LO regenerants to Col-0. **B)** PCAs for each methylation context using the same data are shown in **A**. Original data from Wibowo et al. (2018), ENA accession PRJEB26932.

Hierarchical clustering of methylation rates in 1,282 CHG- and 741 CHH-DMRs revealed tissue type during DNA sampling as the major determinant of methylation within these regions (Figure 4A,B). Tissue of origin was subordinated to sampling tissue type. Principal component analysis (PCA) of methylation rates within the same regions also showed clear ordination based on sampling tissue type along PC1 (Figure 4B), for both CHH and CHG contexts. In contrast, among all 9,487 DMRs identified in CG context, tissue of origin appeared to play a role alongside sampling tissue type. Our analysis also recapitulated the main finding of the original publication, namely that DNA methylation in leaves of RO progenitors was more similar to that in root tissue. Our results generally supported the original conclusion of partial maintenance of epigenetic marks retained from the tissue of origin across generations of sexually reproduced offspring, yet suggested CG methylation to be the main driver of this observation (Figure 4A).

bioRxiv preprint doi: <https://doi.org/10.1101/2022.01.06.475031>; this version posted January 6, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

To validate whether MethylScore could accurately resolve complex data structures in larger plant genomes with much denser DNA methylation profiles, we also re-analysed a regeneration-related dataset from rice (Stroud et al., 2013a). Similar to the *A. thaliana* dataset, MethylScore identified DMRs that separated regenerated from non-regenerated samples (Supplementary Figure 3). Interestingly, however, clustering of samples based on methylation rates within DMRs changed considerably depending on the sequence context. For example, callus samples were similar to control plants and regenerants in CG-DMRs but stood out as hypermethylated in the CHH context (Supplementary Figure 3).

DMR calling on a population scale can identify unknown underlying sample structures

MethylScore does not require information on sample grouping but instead clusters samples for each DMR based on methylation rates, so we expected it to be able to call DMRs even in very large population-scale datasets. We also wanted to explore how combined epigenetic and genetic variation would affect the algorithm. Natural genetic and epigenetic variation in *A. thaliana* has been cataloged in the 1001 Genomes and Epigenomes Project (www.1001genomes.org) (1001 Genomes Consortium, 2016; Kawakatsu et al., 2016). We analyzed a subset of 645 *A. thaliana* accessions that had been sequenced by WGBS at the SALK Institute, San Diego, USA. Applying the population-scale approach of MethylScore, we identified 60,797, 16,627, and 8,406 DMRs in CG, CHG, and CHH, respectively. PCA on the accession-specific methylation rates in these regions revealed geographic clustering that separated a cluster of Central Asian accessions from the rest (Figure 5). This grouping occurred for all three sequence contexts and had not been described in the original publication, indicating that MethylScore's approach can detect data structures that remain hidden with conventional DMR calling tools. Moreover, CG, but not CHG or CHH DMRs, resolved a latitudinal transect for European accessions (Figure 5). We wanted to explore the source of the signal further and used the geographic coordinates of collection sites for each of the *A. thaliana* accessions to retrieve bioclimatic variables from the WorldClim dataset (www.worldclim.org) (Fick and Hijmans, 2017). While the Central Asian cluster showed strong correlation with low annual average temperature (bio1), strongest correlation was observed for the lowest temperature in the coldest month of the year (bio6). While local climate is strongly confounded by geographic location and thus population structure, this observation is in line with previous reports that showed interdependence between DNA methylation and ambient temperature (Dubin et al., 2015) as well as seasonality (Shen et al., 2014).

To further explore MethylScore's capability to resolve hidden sample/population structures, we applied it to WGBS data of 169 epigenetic recombinant inbred lines (epiRILs) (Zhang et al., 2021). These lines are derived from a cross of Col-0 wild-type and a loss-of-function mutant of DDM1 and propagated by single-seed-descent. Genetically, these lines can be considered near-isogenic; previous studies showed that these epiRILs carry marked differences in DNA methylation in mosaic-like patterns of DNA methylation, some of which are stably inherited through generations of single-seed descent (Johannes et al., 2009; Colomé-Tatché et al., 2012). As every epiRIL has been propagated independently and should therefore be independent from every other line, we did not expect any sample relationships. However, when analyzing the DMRs returned by MethylScore's population-scale analysis in a PCA, the epiRILs formed several clusters (Supplementary Figure 4). In all three sequence contexts, most lines were contained in one large cluster, with the exception of a few lines that clustered separately (Supplementary Figure 4). This clustering may reflect the fact that lines were

selected for WGBS based on phenotypic information (i.e., selected epigenotyping) or yet unknown sources of sample stratification, including genetic structure due to shared TE insertion profiles (Quadrana et al., 2019) in a subset of lines.

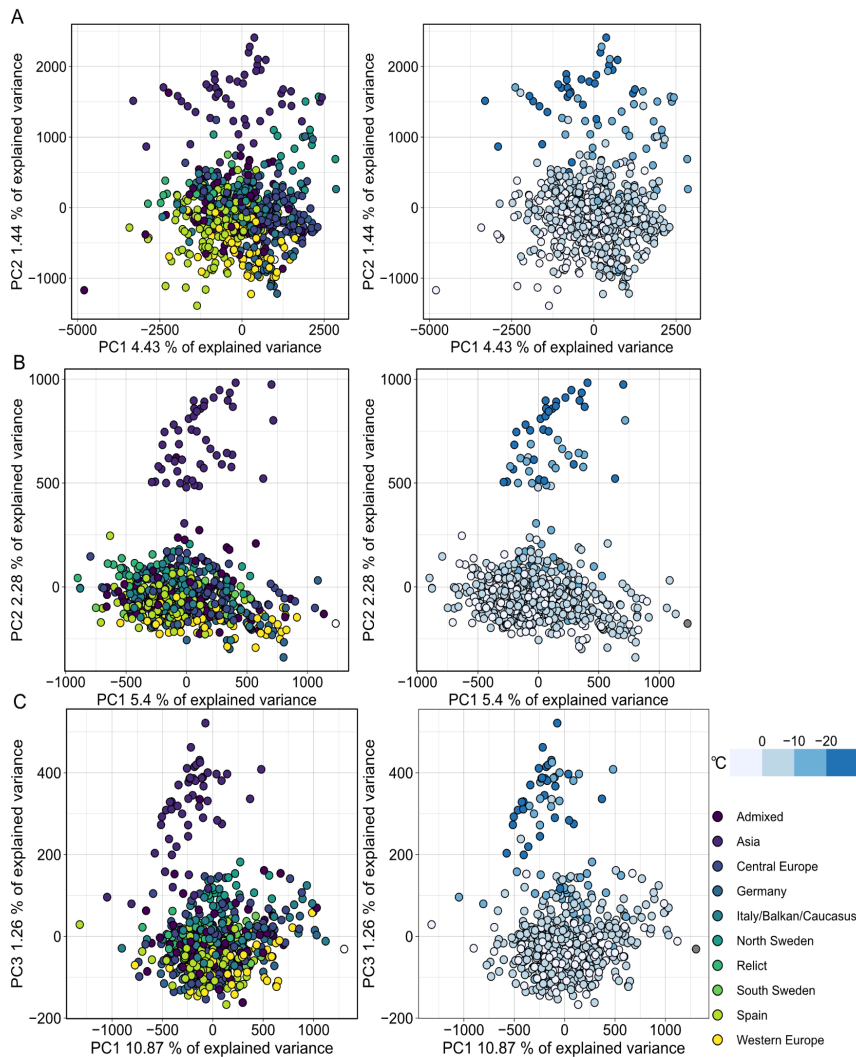


Figure 5: Population structure analysis of natural *A. thaliana* accessions based on DMRs identified by the population-scale clustering approach of MethylScore.

PCA shows group formation in CG (A), CHG (B) and CHH (C) methylation contexts. Colors indicate admixture groups (left column) and seasonality with regard to the lowest temperature in the coldest month (right column). Data was retrieved via geographic coordinates of collection sites for each accession from the worldclim.org bio6 dataset (Fick and Hijmans, 2017). Original WGBS published in Kawakatsu et al. (2016).

bioRxiv preprint doi: <https://doi.org/10.1101/2022.01.06.475031>; this version posted January 6, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

MethylScore DMRs reveal recurrent trans-acting association signals in A. thaliana natural accessions

MethylScore's *k*-means clustering leads to grouping of samples into only a small number of groups (mostly two or three), hence within-group variance becomes relatively large when sample numbers are high, as was the case for the 1001 Genomes data. As a result, DMRs with noticeable between-group differences become rare, explaining the relatively low number of DMRs we observed in the 1001 Genomes data, in relation to the size of the dataset. To increase the statistical power to detect genotype-epigenotype associations in a genome-wide association (GWA) mapping approach that uses methylation rates in each DMR as the phenotype vector, we changed from a population-scale to a pairwise DMR calling, comparing each of the 645 accessions individually to the Col-0 reference. On average, each pairwise comparison identified 1,708, 778, and 3,625 DMRs in CG, CHG, and CHH contexts, respectively. Many DMRs from different pairwise comparisons partially or fully overlapped, so we created sets of unions of overlapping DMRs for each context, resulting in 18,044 CG-, 9,674 CHG-, and 25,350 CHH-DMRs.

We queried methylation rates per accession and conducted GWAS analyses, using region-level methylation averages as the phenotype vector. On the genotype level, we considered all genotyped SNPs among the 645 accessions that had a minor allele frequency (MAF) > 5% (1,813,837 SNPs). Using this strategy, we identified 16,083 regions in CG, 8,102 in CHG, and 19,388 in CHH context that yielded at least one SNP association with significance levels that passed the Bonferroni-corrected significance threshold at $\alpha = 0.05$ (threshold = 2.8×10^{-8}).

Combining these results in a meta-analysis revealed that many of these associations marked region-proximal SNP locations in *cis* for CG (Supplementary Figure 5), CHG and CHH (Figure 6) methylation contexts. Intriguingly, a few also tagged several recurring association signals in *trans* (Figure 6). For CHG DMRs, we found two recurrent associations with markers on chromosomes 3 and 5, respectively (Figure 6A,D,E). Chr3:4,496,047 was found to be associated with mean methylation rates in 97 DMRs with consistently decreased methylation in accessions carrying the alternative allele (Figure 6F,G). The SNP was located in close proximity to the *miR823A* locus, which encodes the primary miRNA (pri-miRNA) of microRNA (miRNA) 823A, known to target *CMT3*. *CMT3* is the main methyltransferase depositing CHG methylation (Lindroth et al., 2001) and via its chromodomain provides a direct link between DNA methylation and H3K9 dimethylation at constitutive heterochromatin in a concerted feedback loop with KRYPTONITE (KYP) and SUPPRESSOR OF VARIATION 3-9 HOMOLOGUE PROTEIN 5/6 (*SUVH5/6*) (Du et al., 2012; Jackson et al., 2002; Du et al., 2014; Ebbs and Bender, 2006).

The second CHG-DMR-associated locus, Chr5:23,553,506, showed association in 44 DMRs, similarly with relatively lower methylation levels in accessions carrying the alternative allele. The SNP resolved a genomic region upstream of *MULTICOPY SUPPRESSOR OF IRA1* (*MSI1*) and appeared to be in linkage with several markers in that region (Figure 6D). *MSI1* acts in the evolutionarily conserved Retinoblastoma pathway and is implicated in genomic imprinting of the *FLOWERING WAGENINGEN* (*FWA*) and *FERTILIZATION INDEPENDENT SEED 2* (*FIS2*) in *A. thaliana* via direct interaction with *RETINOBLASTOMA RELATED 1* (*RBR1*) protein (Jullien et al., 2008). Interestingly, both *RBR1* and *MSI1* have also been identified as integral components of the *Arabidopsis* DREAM complex (Ning et al., 2020).

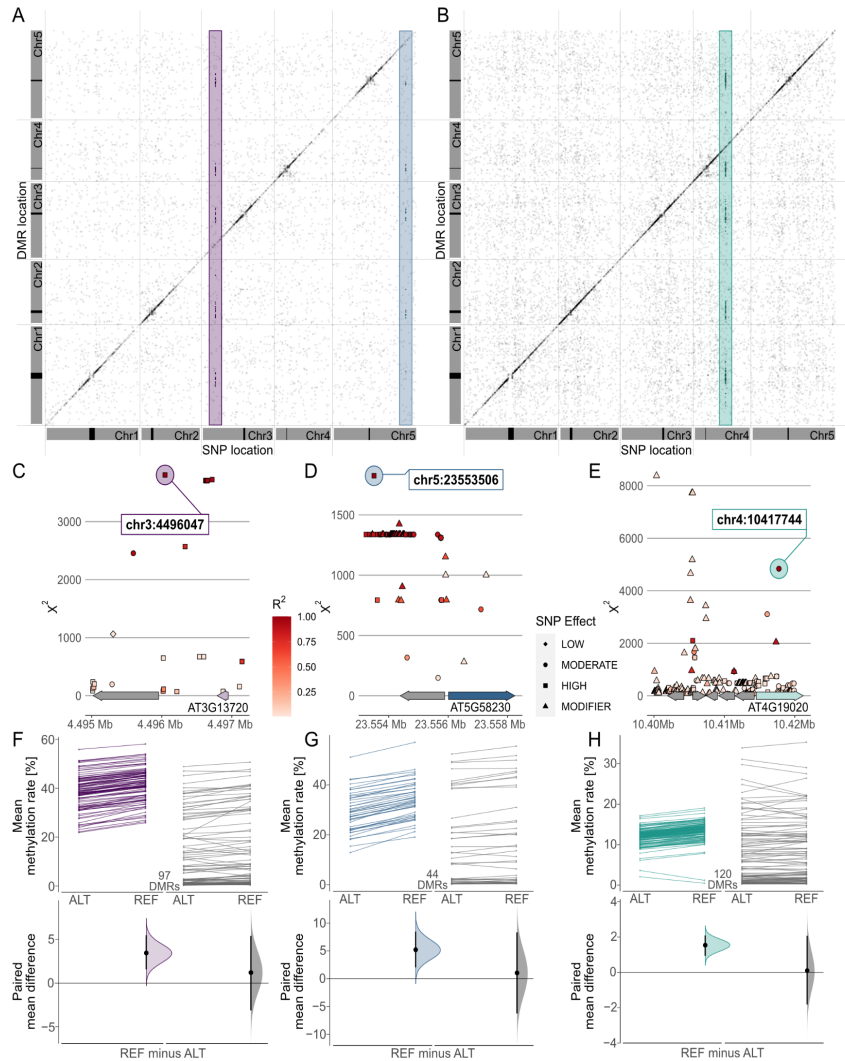


Figure 6: Genome Wide Association (GWA) signals recurrently emerge from differential methylation found across the *A. thaliana* 1001 Methylomes panel (Kawakatsu et al., 2016).

GWA analyses on region-level methylation rate averages reveal recurrent signals in CHG (**A**) and CHH (**B**) methylation contexts. For each DMR, only top ranked SNPs that pass the Bonferroni corrected significance threshold at $\alpha = 0.05$ are included, based on the number of SNP markers available across all 646 *A. thaliana* accessions used in the study (1,813,837 SNPs with minor allele frequency (MAF) > 5%, $p < 2.8 \times 10^{-8}$). **C-E**) Genomic loci of recurrent trans-acting SNPs highlighted in (**A**) and (**B**). **F-H**) Effect sizes of SNPs highlighted in (**A**) and (**B**) on methylation rates in regions underlying the SNP association are shown as slopegraphs and bootstrap estimates for carriers of the alternative (ALT) and reference (REF) alleles, respectively. In each plot, an equally sized set of randomly selected DMRs (in gray) is included for comparison.

bioRxiv preprint doi: <https://doi.org/10.1101/2022.01.06.475031>; this version posted January 6, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

In the CHH context, our finding of a recurrent association at Chr5:10417744 (Figure 6B,E) confirms a recent study that identified natural alleles of *CHROMOMETHYLASE2* (*CMT2*) as determinants of CHH methylation (Sasaki et al., 2019). The association signal emerged from 120 DMRs, showing a general trend of lower methylation in accessions carrying the alternative allele, with the exception of two regions of overall low CHH methylation (Figure 6H). In summary, this shows that MethylScore provides context-specific DMRs from large WGBS datasets that can be used to determine specific genome-epigenotype associations.

Discussion

Genome-wide studies of DNA methylation typically aim to reveal differences in DNA methylation between samples or groups of samples. This includes identification of phenotypically relevant epialleles, finding epigenetically controlled regulatory loci, or assessing naturally occurring epigenetic variation. Here, we have presented MethylScore, a pipeline for the identification and characterization of differentially methylated loci specifically from plant WGBS data. We built MethylScore in such a way as to make it accessible to a broad community of plant researchers, requiring minimal computational background; provided with alignment files or methylation metrics files and a simple sample information table, MethylScore can be run with a single command line.

Currently available DMR callers are mostly built on assumptions based on mammalian DNA methylation, including the sequence context, genomic distribution and frequency of methylated cytosines, and the average methylation rate. Plant DNA methylation violates many of these assumptions. Moreover, most DMR callers typically follow one of three general strategies, DMP clustering, genome tiling, or pre-defining regions of interest, each with their respective caveats. DMP clustering first tests for DMPs and merges spatially adjacent loci to define DMRs. Examples of such tools are *DSS* (Feng et al., 2014), *Bisulfighter* (Saito et al., 2014), *MOABS* (Sun et al., 2014), *BSmooth* (Hansen et al., 2012), and *methylypy* (Ziller et al., 2013; Schultz et al., 2015). DMP analysis typically involves statistical tests on each cytosine in the genome, racking up a heavy multiple testing burden. If no appropriate region-level testing is applied, this multiple testing problem is carried over to the DMRs, resulting in inappropriate false discovery rate (FDR) control caused by significance thresholding at single loci (Korthauer et al., 2018). In addition, many such approaches classify DMRs based on mere presence of DMPs without considering the directionality of methylation change. Multiple-testing limitations also apply to tiling approaches in which testable segments correspond to windows or sliding windows along the genome, implemented for example in *methylyKit* (Akalin et al., 2012). The third category of DMR callers use predefined regions, selected based on existing knowledge, e.g. genome annotation features. These can be gene bodies, promoter regions, transposons, or predetermined CpG islands. While this reduces the multiple-testing problem, it ignores potentially relevant loci that are not included in the predefined set, yet makes tools such as *BiSeq* (Hebestreit et al., 2013) suited for targeted sequencing approaches such as reduced-representation bisulfite sequencing (RRBS) (Meissner et al., 2005).

All of the above methods have additional downsides when applied to *plant* WGBS data. Methylation occurring in all sequence contexts in plants aggravates the DMP multiple testing problem. Some approaches, including *BSmooth* (Hansen et al., 2012), *BiSeq* (Hebestreit et al., 2013), and *dmrseq* (Korthauer et al., 2018), moreover assume local correlation between spatially adjacent methylated cytosines to compute smoothed methylation estimates. Despite

the presence of CHH islands that have been reported for some species (Zemach et al., 2010; Gent et al., 2013), this assumption primarily holds true for studies in mammalian genomes with high density of methylated CpG regions. It is, however, less applicable in plants and potentially affects the performance of these methods in determining genomic boundaries of methylated regions. Similarly, window- or tile-based approaches are best applicable when methylation is evenly and densely distributed. However, in *A. thaliana* and other plant genomes, the overwhelming share of methylated cytosines are concentrated in only a small fraction of the genome. Window-based approaches therefore cause a substantial penalty in statistical power when controlling for FDR, because many of the statistical tests lack biological relevance in absence of methylation in the windows that are tested.

MethylScore does not require a priori information on DNA methylation parameters or dataset structures

In an unsupervised training step, MethylScore identifies the methylation rate distribution per sequence context in the actual data, followed by classifying the genome into states of high and low methylation based on these training parameters. Alternatively, users can decide to train MethylScore on a reference sample and apply the training parameters on all other samples in their dataset. Only those regions are tested for differential methylation in which high-methylation states differ across samples, and so our pipeline avoids testing regions with no or very little DNA methylation variation across samples, which drastically reduces the number of necessary statistical tests compared to other methods.

As shown in the test cases above, MethylScore can handle small datasets with only few samples to very large datasets with hundreds of WGBS runs. Unless users decide to provide information on replicate groups, the pipeline applies a population-scale analysis, clustering samples into groups with similar methylation rates for each testable genomic segment. This has two effects: first, MethylScore is thus able to reveal hidden sample structures that researchers might not be aware of at the start of the study. Second, substantial within-group variance, e.g. within a group of replicates of the same treatment condition, will become apparent in the DMR output and will not be masked by a forced *a priori* grouping of replicates.

A modular pipeline that can be integrated with other WGBS analysis pipelines

MethylScore is designed as a post-alignment, secondary analysis workflow starting from either alignments in *bam* format or pre-existing single-cytosine metrics in *bedGraph* format, in contrast to existing end-to-end solutions such as *wg-blimp* (Wöste et al., 2020) or *PiGx bsseq* (Wurmus et al., 2018). However, its modular design also allows for tight integration with primary analysis pipelines such as *nf-core methylseq* (Ewels et al., 2020), *snakePipes WGBS* (Bhardwaj et al., 2019), or *EpiDiverse Toolkit* (Nunn et al., 2021). These pipelines take raw BS-sequencing reads as an input and perform extensive quality control, along with trimming and read mapping, which is highly advised to rule out or mitigate potential biases arising from experimental factors such as library preparation protocols, adapter content, and sequence duplication levels, before attempting to identify DMRs.

bioRxiv preprint doi: <https://doi.org/10.1101/2022.01.06.475031>; this version posted January 6, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Limitations and shortcomings of the pipeline

MethylScore is extensively parameterised. While we show that the default parameter set is suitable on data derived from *A. thaliana* and *O. sativa*, analysis of other species with different methylation characteristics might require a parameter sweep for optimal results.

Outlook

Despite recent advances in detecting chemically modified bases in long-read sequencing data, WGBS remains the current standard for the analysis of whole-genome DNA methylation profiles. This applies even more to plants, for which the complexity and diversity of base modifications poses major hurdles to long-read-based technologies. Even if these technical issues will be overcome in future, statistical analysis of differential methylation to determine relevant epiallelic loci will remain a key challenge in this type of studies.

Materials and methods

Pipeline strategy

When provided with alignments as input, MethylScore can first merge and coordinate sort technical replicates using *samtools* v1.9 (Li et al., 2009) and subsequently deduplicates reads arisen from PCR amplification using *picard* v2.26.8 (Picard toolkit, 2019). To enable parallel processing, alignments are split by chromosome to concurrently extract single cytosine metrics using *MethylDackel* v0.4.0 (<https://github.com/dpryan79/MethylDackel>). The resulting read counts are collected for all samples and tabulated into an intermediate, custom-format table containing chromosomal position, sequence context, strand information, unmethylated and methylated read counts, as well as derived metrics such as coverage and methylation rate per cytosine. This information is collected in the genome matrix, a single file containing all of the above information for every sample and for each cytosine covered in at least one sample.

Using this genome-wide information, a two-state HMM is trained iteratively using the Baum-Welch algorithm for each of the samples as described by (Molaro et al., 2011) with two key adaptations. First, to account for the fact that, in contrast to DNA methylation in mammals, plant genomes often only harbor DNA methylation in a small fraction of the genome, the original implementation was adapted to detect *hyper*-methylated rather than *hypo*-methylated regions, effectively achieved by inverting the methylation rates. Second, we adapted the extension to a total of three sequence contexts, allowing the model to learn distinct parameters independently for CG, CHG and CHH sequence contexts for both states. This results in a total of six distributions with corresponding emission probabilities as well as transition probabilities between the methylated and unmethylated states.

Given the parameter estimates after model convergence for a maximum of 30 iterations, the most likely path is determined by Posterior Decoding and used for probabilistic segmentation of the genome into methylated regions (MRs). During this step, MR breakpoints are introduced at either chromosome boundaries or when a parameter is exceeded that defines the maximum sequence length devoid of covered cytosine positions (named "desert size", default value 100 bp). This user-defined parameter prevents extending MRs over 100 bp without any methylation

information. Having segmented the genome into MRs, MethylScore then applies two complementary population-scale approaches to reduce complexity of the statistical testing framework. First, candidate regions for significance testing are restricted to those with a user-defined change in MR frequency across samples (default: 20% change of MR frequency within 30 bp), indicative of natural region boundaries that are widespread in the sample population. Additional filters can be set to avoid testing short or unreliable MRs by requiring a minimum number of covered cytosines (default: 5) and a lower bound for coverage at those sites (default: 3). The second approach used by MethylScore to further decrease the number of statistical tests is to employ k -means clustering in order to group samples by mean methylation rates within any given MR. For each candidate region, cluster centers are searched to minimize the within-group variance. The value of k is iteratively incremented starting from $k = 2$, until the pairwise comparison of all cluster centers results in a methylation difference of less than a user-defined value, in which case the previous k is chosen. This parameter lets the user choose a target minimum methylation difference between groups of samples and thus can be effectively used to discard likely biologically irrelevant DMRs with only a few percentage points methylation difference.

Remaining regions that fulfill aforementioned criteria are tested for differential methylation using a beta-binomial-based test; resulting p-values are subsequently corrected for multiple tests as described in (Becker et al., 2011).

Implementation in Nextflow

To achieve high standards of reproducibility and portability of pipeline execution across different compute infrastructures, the workflow manager Nextflow (Di Tommaso et al., 2017) was chosen to implement MethylScore. Nextflow enables containerised execution using a prebuilt Docker image (available at <https://quay.io/repository/beckerlab/methylscore>), that contains all necessary dependencies including perl libraries, the pre-compiled HMM implementation, and version-pinned bioconda packages. Furthermore, Nextflow ensures efficient use of available computational infrastructure by straightforward parallelisation of tasks, following a scatter-gather strategy.

Re-analysis of published datasets

WGBS reads were retrieved from the ENA and GEO repositories PRJEB26932 (Wibowo et al., 2016), PRJEB12413 (Tedeschi et al., 2019), GSE137754 (Ning et al., 2020) GSE42410 (Stroud et al., 2013a) and [GSE171414](#) (Zhang et al., 2021)

Raw data were trimmed and mapped to the TAIR10 (Lamesch et al., 2012) reference genome for *A. thaliana* data and the IGRSP-1.0 genome (Kawahara et al., 2013) for *O. sativa* spp. *japonica* data using v1.6 of the *nf-core methylseq* pipeline (Ewels et al., 2021). Standard settings for Illumina adapter trimming, and the default aligner *bismark* (Krueger and Andrews, 2011) were used with addition of the `--relax_mismatches` preset. The rich quality control output of the pipeline was used to assess methylation bias on both read ends and affected bases were ignored for downstream analysis with MethylScore.

1001 Methylomes analysis

Single-cytosine metrics obtained from the *A. thaliana* 1001 Methylomes dataset (Kawakatsu et al., 2016) were used to call DMRs in pairwise comparisons between 645 accessions using

bioRxiv preprint doi: <https://doi.org/10.1101/2022.01.06.475031>; this version posted January 6, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Col-0 as the point of reference. Overlapping regions found in multiple comparisons were merged using the *plyranges* (v.1.14.0 (Lee et al., 2019) *reduce_ranges()* function in *R* (v4.1.2 (R Core Team, 2021) For each sequence context, methylation rates within all identified DMRs were retrieved and used to seek associations between the average methylation rate for each accession within each DMR and the underlying genotype by conducting Genome-Wide Association Studies (GWAS) using the *Limix* framework (Lippert et al., 2014) for linear mixed models using an in-house python implementation packaged into a containerised Nextflow pipeline (<https://gitlab.lrz.de/beckerlab/gwas-nf>).

In brief, our GWAS pipeline retrieves the average methylation rate for each accession within a given DMR and estimates population structure from all SNPs using the *limix.stats.linear_kinship* function in *Limix*. GWAS is performed via the *limix.qtl.scan* function after subsetting the full genome SNP matrix (1001 Genomes Consortium, 2016) to include all polymorphic SNPs that have a minor allele frequency (MAF) of at least 5% across the population of 645 accessions. For meta analyses, only those SNP positions are reported for which the significance level exceeds the Bonferroni corrected threshold with respect to the total number of included SNP markers (1,813,837).

Data visualization

Figures were generated using *R* 4.1.2 (R Core Team, 2021) and *Bioconductor* 3.14 (Gentleman et al., 2004; Huber et al., 2015). Heatmaps of methylation rates were generated using the *ComplexHeatmap* library (v2.10.0 (Gu et al., 2016)). Principal components were computed using the *pcaMethods* library (v1.86.0 (Stacklies et al., 2007) and visualized with *ggplot2* (v3.3.5 (Wickham, 2009) and the *ggrepel* library (v0.9.1 (Slowikowski)). SNP linkage was calculated using the *snpStats* library (v1.44.0) and genomic locations were illustrated using the *gggenes* library (v0.4.1 (Wilkins)). Genome wide representations of SNP associations were drawn using *ggplot2* (v3.3.5 (Wickham, 2009) and the *ggforce* (v0.3.3 (Pedersen) library for visual annotation. Slopegraphs and Bootstrap Estimation plots were generated using the *dabestr* library (v0.3.0 (Ho et al., 2019)).

Acknowledgements

We would like to thank Niklas Schandry, Eva Knoch, and Daniela Ramos Cruz for critical reading of the manuscript and helpful suggestions. The computational results presented were obtained using the CLIP cluster at Vienna BioCenter (VBC) and the BioHPC Genomics cluster housed at the Leibniz Supercomputing Centre (LRZ).

This research was supported by the Austrian Academy of Sciences (P.H., C.B., R.P.); the European Union's Horizon 2020 research and innovation programme via the European Research Council (ERC) Grant agreement No. 716823 "FEAR-SAP" (P.H., C.B.), DFG project ERA-CAPS AUREATE (D.W.), and via the Marie Skłodowska-Curie ETN 'EpiDiverse' (A.N., C.B.), grant agreement no. 764965 "Epidiverse". F.J. acknowledges support from the Technical University of Munich-Institute for Advanced Study funded by the German Excellent Initiative and the European Seventh Framework Program under grant agreement no. 291763. F.J., and I.K. were supported by the SFB Sonderforschungsbereich924 of the Deutsche Forschungsgemeinschaft (DFG).

Author contributions

P.H., J.H., F.J., D.W., and C.B. conceived the study and developed the pipeline; D.L. and C.B. supervised the work. J.H. wrote the MethylScore code; P.H. built the Nextflow pipeline. I.K. generated the epiRIL data. P.H. and A.N. analyzed the data. R.P. prepared methylation calls on the *A. thaliana* 1001 Genomes dataset. P.H. and C.B. wrote the manuscript.

Competing Interests

The authors declare the following competing interests: JH is currently an employee of Computomics GmbH. S.J.S. is currently the CEO of and holds shares in Computomics GmbH.

Data availability

The MethylScore pipeline is available on Github: <https://github.com/Computomics/MethylScore>. The pipeline that was used to conduct GWA studies is available on Gitlab: <https://gitlab.lrz.de/beckerlab/gwas-nf>.

References

- 1001 Genomes Consortium** (2016). 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell* **166**: 481–491.
- Akalin, A., Kormaksson, M., Li, S., Garrett-Bakelman, F.E., Figueroa, M.E., Melnick, A., and Mason, C.E.** (2012). methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol.* **13**: R87.
- Aufsatz, W., Florian Mette, M., van der Winden, J., Matzke, A.J.M., and Matzke, M.** (2002). RNA-directed DNA methylation in *Arabidopsis*. *Proc. Natl. Acad. Sci. U. S. A.* **99**: 16499–16506.
- Becker, C., Hagmann, J., Müller, J., Koenig, D., Stegle, O., Borgwardt, K., and Weigel, D.** (2011). Spontaneous epigenetic variation in the *Arabidopsis thaliana* methylome. *Nature* **480**: 245–249.
- Bhardwaj, V., Heyne, S., Sikora, K., Rabbani, L., Rauer, M., Kilpert, F., Richter, A.S., Ryan, D.P., and Manke, T.** (2019). snakePipes: facilitating flexible, scalable and integrative epigenomic analysis. *Bioinformatics* **35**: 4757–4759.
- Cao, X., Aufsatz, W., Zilberman, D., Mette, M.F., Huang, M.S., Matzke, M., and Jacobsen, S.E.** (2003). Role of the DRM and CMT3 methyltransferases in RNA-directed DNA methylation. *Curr. Biol.* **13**: 2212–2217.
- Cao, X. and Jacobsen, S.E.** (2002). Role of the arabidopsis DRM methyltransferases in de novo DNA methylation and gene silencing. *Curr. Biol.* **12**: 1138–1144.
- Cervera, M.T., Ruiz-García, L., and Martínez-Zapater, J.M.** (2002). Analysis of DNA methylation in *Arabidopsis thaliana* based on methylation-sensitive AFLP markers. *Mol. Genet. Genomics* **268**: 543–552.
- Cokus, S.J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C.D., Pradhan, S., Nelson, S.F., Pellegrini, M., and Jacobsen, S.E.** (2008). Shotgun bisulphite

bioRxiv preprint doi: <https://doi.org/10.1101/2022.01.06.475031>; this version posted January 6, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

- sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature* **452**: 215–219.
- Colomé-Tatché, M. et al.** (2012). Features of the Arabidopsis recombination landscape resulting from the combined loss of sequence variation and DNA methylation. *Proc. Natl. Acad. Sci. U. S. A.* **109**: 16240–16245.
- Cubas, P., Vincent, C., and Coen, E.** (1999). An epigenetic mutation responsible for natural variation in floral symmetry. *Nature* **401**: 157–161.
- Deaton, A.M. and Bird, A.** (2011). CpG islands and the regulation of transcription. *Genes Dev.* **25**: 1010–1022.
- Di Tommaso, P., Chatzou, M., Floden, E.W., Barja, P.P., Palumbo, E., and Notredame, C.** (2017). Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* **35**: 316–319.
- Dubin, M.J. et al.** (2015). DNA methylation in Arabidopsis has a genetic basis and shows evidence of local adaptation. *Elife* **4**: e05255.
- Du, J. et al.** (2012). Dual binding of chromomethylase domains to H3K9me2-containing nucleosomes directs DNA methylation in plants. *Cell* **151**: 167–180.
- Du, J., Johnson, L.M., Groth, M., Feng, S., Hale, C.J., Li, S., Vashisht, A.A., Wohlschlegel, J.A., Patel, D.J., and Jacobsen, S.E.** (2014). Mechanism of DNA methylation-directed histone methylation by KRYPTONITE. *Mol. Cell* **55**: 495–504.
- Ebbs, M.L. and Bender, J.** (2006). Locus-specific control of DNA methylation by the Arabidopsis SUVH5 histone methyltransferase. *Plant Cell* **18**: 1166–1176.
- Eichten, S.R. et al.** (2011). Heritable epigenetic variation among maize inbreds. *PLoS Genet.* **7**: e1002372.
- Ewels, P. et al.** (2021). nf-core/methylseq: nf-core/methylseq version 1.6.1 [Nauseous Serpent].
- Ewels, P.A., Peltzer, A., Fillinger, S., Patel, H., Alneberg, J., Wilm, A., Garcia, M.U., Di Tommaso, P., and Nahnsen, S.** (2020). The nf-core framework for community-curated bioinformatics pipelines. *Nat. Biotechnol.* **38**: 276–278.
- Feng, H., Conneely, K.N., and Wu, H.** (2014). A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. *Nucleic Acids Res.* **42**: e69.
- Fick, S.E. and Hijmans, R.J.** (2017). WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *Int. J. Climatol.* **37**: 4302–4315.
- Finnegan, E.J., Peacock, W.J., and Dennis, E.S.** (1996). Reduced DNA methylation in *Arabidopsis thaliana* results in abnormal plant development. *Proc. Natl. Acad. Sci. U. S. A.* **93**: 8449–8454.
- Frommer, M., McDonald, L.E., Millar, D.S., Collis, C.M., Watt, F., Grigg, G.W., Molloy, P.L., and Paul, C.L.** (1992). A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc. Natl. Acad. Sci. U. S. A.* **89**: 1827–1831.
- Gehring, M., Huh, J.H., Hsieh, T.-F., Penterman, J., Choi, Y., Harada, J.J., Goldberg,**

- R.B., and Fischer, R.L.** (2006). DEMETER DNA glycosylase establishes MEDEA polycomb gene self-imprinting by allele-specific demethylation. *Cell* **124**: 495–506.
- Gent, J.I., Ellis, N.A., Guo, L., Harkess, A.E., Yao, Y., Zhang, X., and Dawe, R.K.** (2013). CHH islands: de novo DNA methylation in near-gene chromatin regulation in maize. *Genome Res.* **23**: 628–637.
- Gentleman, R.C. et al.** (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5**: R80.
- van der Graaf, A., Wardenaar, R., Neumann, D.A., Taudt, A., Shaw, R.G., Jansen, R.C., Schmitz, R.J., Colomé-Tatché, M., and Johannes, F.** (2015). Rate, spectrum, and evolutionary dynamics of spontaneous epimutations. *Proc. Natl. Acad. Sci. U. S. A.* **112**: 6676–6681.
- Gu, Z., Eils, R., and Schlesner, M.** (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**: 2847–2849.
- Hagmann, J., Becker, C., Müller, J., Stegle, O., Meyer, R.C., Wang, G., Schneeberger, K., Fitz, J., Altmann, T., Bergelson, J., Borgwardt, K., and Weigel, D.** (2015). Century-scale methylome stability in a recently diverged *Arabidopsis thaliana* lineage. *PLoS Genet.* **11**: e1004920.
- Hansen, K.D., Langmead, B., and Irizarry, R.A.** (2012). BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol.* **13**: 1–10.
- Hebestreit, K., Dugas, M., and Klein, H.-U.** (2013). Detection of significantly differentially methylated regions in targeted bisulfite sequencing data. *Bioinformatics* **29**: 1647–1653.
- Henderson, I.R. and Jacobsen, S.E.** (2007). Epigenetic inheritance in plants. *Nature* **447**: 418–424.
- Heyn, H., Moran, S., Hernando-Herraez, I., Sayols, S., Gomez, A., Sandoval, J., Monk, D., Hata, K., Marques-Bonet, T., Wang, L., and Esteller, M.** (2013). DNA methylation contributes to natural human variation. *Genome Res.* **23**: 1363–1372.
- Ho, J., Tumkaya, T., Aryal, S., Choi, H., and Claridge-Chang, A.** (2019). Moving beyond P values: data analysis with estimation graphics. *Nat. Methods* **16**: 565–566.
- Huber, W. et al.** (2015). Orchestrating high-throughput genomic analysis with Bioconductor. *Nat. Methods* **12**: 115–121.
- Huff, J.T. and Zilberman, D.** (2014). Dnmt1-independent CG methylation contributes to nucleosome positioning in diverse eukaryotes. *Cell* **156**: 1286–1297.
- Jackson, J.P., Lindroth, A.M., Cao, X., and Jacobsen, S.E.** (2002). Control of CpNpG DNA methylation by the KRYPTONITE histone H3 methyltransferase. *Nature* **416**: 556–560.
- Jaenisch, R. and Bird, A.** (2003). Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat. Genet.* **33**: 245–254.
- Jeltsch, A.** (2006). On the enzymatic properties of Dnmt1: specificity, processivity, mechanism of linear diffusion and allosteric regulation of the enzyme. *Epigenetics* **1**: 63–66.
- Johannes, F. et al.** (2009). Assessing the impact of transgenerational epigenetic variation

bioRxiv preprint doi: <https://doi.org/10.1101/2022.01.06.475031>; this version posted January 6, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

on complex traits. *PLoS Genet.* **5**: e1000530.

Jühling, F., Kretzmer, H., Bernhart, S.H., Otto, C., Stadler, P.F., and Hoffmann, S. (2016). metilene: fast and sensitive calling of differentially methylated regions from bisulfite sequencing data. *Genome Res.* **26**: 256–262.

Jullien, P.E., Mosquna, A., Ingouff, M., Sakata, T., Ohad, N., and Berger, F. (2008). Retinoblastoma and its binding partner MSI1 control imprinting in Arabidopsis. *PLoS Biol.* **6**: e194.

Kankel, M.W., Ramsey, D.E., Stokes, T.L., Flowers, S.K., Haag, J.R., Jeddloh, J.A., Riddle, N.C., Verbsky, M.L., and Richards, E.J. (2003). Arabidopsis MET1 cytosine methyltransferase mutants. *Genetics* **163**: 1109–1122.

Kawahara, Y. et al. (2013). Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. *Rice* **6**: 4.

Kawakatsu, T. et al. (2016). Epigenomic diversity in a global collection of *Arabidopsis thaliana* accessions. *Cell* **166**: 492–505.

Korthauer, K., Chakraborty, S., Benjamini, Y., and Irizarry, R.A. (2018). Detection and accurate false discovery rate control of differentially methylated regions from whole genome bisulfite sequencing. *Biostatistics* **20**: 367–383.

Krueger, F. and Andrews, S.R. (2011). Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**: 1571–1572.

Lamesch, P. et al. (2012). The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.* **40**: D1202–10.

Law, J.A. and Jacobsen, S.E. (2010). Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat. Rev. Genet.* **11**: 204–220.

Lee, S., Cook, D., and Lawrence, M. (2019). plyranges: a grammar of genomic data transformation. *Genome Biol.* **20**: 4.

Liégard, B., Baillet, V., Etcheverry, M., Joseph, E., Lariagon, C., Lemoine, J., Evrard, A., Colot, V., Gravot, A., Manzanares-Dauleux, M.J., and Jubault, M. (2019). Quantitative resistance to clubroot infection mediated by transgenerational epigenetic variation in Arabidopsis. *New Phytol.* **222**: 468–479.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

Lindroth, A.M., Cao, X., Jackson, J.P., Zilberman, D., McCallum, C.M., Henikoff, S., and Jacobsen, S.E. (2001). Requirement of CHROMOMETHYLASE3 for maintenance of CpXpG methylation. *Science* **292**: 2077–2080.

Lippert, C., Casale, F.P., Rakitsch, B., and Stegle, O. (2014). LIMIX: genetic analysis of multiple traits. bioRxiv: 003905.

Lippman, Z. et al. (2004). Role of transposable elements in heterochromatin and epigenetic control. *Nature* **430**: 471–476.

Lister, R., O'Malley, R.C., Tonti-Filippini, J., Gregory, B.D., Berry, C.C., Harvey Millar,

- A., and Ecker, J.R.** (2008). Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell* **133**: 523–536.
- Liu, G., Xia, Y., Liu, T., Dai, S., and Hou, X.** (2018). The DNA methylome and association of differentially methylated regions with differential gene expression during heat stress in *Brassica rapa*. *Int. J. Mol. Sci.* **19**.
- MacQueen, J.** (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics* (University of California Press: Berkeley, CA), pp. 281–298.
- Manning, K., Tör, M., Poole, M., Hong, Y., Thompson, A.J., King, G.J., Giovannoni, J.J., and Seymour, G.B.** (2006). A naturally occurring epigenetic mutation in a gene encoding an SBP-box transcription factor inhibits tomato fruit ripening. *Nat. Genet.* **38**: 948–952.
- Manolio, T.A. et al.** (2009). Finding the missing heritability of complex diseases. *Nature* **461**: 747–753.
- Matzke, M.A. and Mosher, R.A.** (2014). RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. *Nat. Rev. Genet.* **15**: 394–408.
- Meissner, A., Gnirke, A., Bell, G.W., Ramsahoye, B., Lander, E.S., and Jaenisch, R.** (2005). Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res.* **33**: 5868–5877.
- Miura, A., Yonebayashi, S., Watanabe, K., Toyama, T., Shimada, H., and Kakutani, T.** (2001). Mobilization of transposons by a mutation abolishing full DNA methylation in Arabidopsis. *Nature* **411**: 212–214.
- Miura, K., Agetsuma, M., Kitano, H., Yoshimura, A., Matsuoka, M., Jacobsen, S.E., and Ashikari, M.** (2009). A metastable DWARF1 epigenetic mutant affecting plant stature in rice. *Proc. Natl. Acad. Sci. U. S. A.* **106**: 11218–11223.
- Molaro, A., Hodges, E., Fang, F., Song, Q., McCombie, W.R., Hannon, G.J., and Smith, A.D.** (2011). Sperm methylation profiles reveal features of epigenetic inheritance and evolution in primates. *Cell* **146**: 1029–1041.
- Niederhuth, C.E. et al.** (2016). Widespread natural variation of DNA methylation within angiosperms. *Genome Biol.* **17**: 194.
- Ning, Y.-Q., Liu, N., Lan, K.-K., Su, Y.-N., Li, L., Chen, S., and He, X.-J.** (2020). DREAM complex suppresses DNA methylation maintenance genes and precludes DNA hypermethylation. *Nat Plants* **6**: 942–956.
- Nunn, A., Can, S.N., Otto, C., Fasold, M., Díez Rodríguez, B., Fernández-Pozo, N., Rensing, S.A., Stadler, P.F., and Langenberger, D.** (2021). EpiDiverse Toolkit: a pipeline suite for the analysis of bisulfite sequencing data in ecological plant epigenetics. *NAR Genom Bioinform* **3**: lqab106.
- Ong-Abdullah, M. et al.** (2015). Loss of Karma transposon methylation underlies the mantled somaclonal variant of oil palm. *Nature* **525**: 533–537.
- Papareddy, R.K., Páldi, K., Smolka, A.D., Hüther, P., Becker, C., and Nodine, M.D.** (2021). Repression of CHROMOMETHYLASE 3 prevents epigenetic collateral damage

bioRxiv preprint doi: <https://doi.org/10.1101/2022.01.06.475031>; this version posted January 6, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

- in *Arabidopsis*. *eLife* **10**: e69396.
- Pedersen, T.L.** ggforce: Accelerating ggplot2 (Github).
- Picard toolkit** (2019). Broad Institute, GitHub repository.
- Pignatta, D., Novitzky, K., Satyaki, P.R.V., and Gehring, M.** (2018). A variably imprinted epiallele impacts seed development. *PLoS Genet.* **14**: e1007469.
- Quadrana, L., Etcheverry, M., Gilly, A., Caillieux, E., Madoui, M.-A., Guy, J., Bortolini, A., Engelen, S., Baillet, V., Wincker, P., Aury, J.-M., and Colot, V.** (2019). Transposition favors the generation of large effect mutations that may facilitate rapid adaption. *Nat. Commun.* **10**: 3421.
- R Core Team** (2021). R: A Language and Environment for Statistical Computing.
- Riddle, N.C. and Richards, E.J.** (2002). The control of natural variation in cytosine methylation in *Arabidopsis*. *Genetics* **162**: 355–363.
- Ronemus, M.J., Galbiati, M., Ticknor, C., Chen, J., and Dellaporta, S.L.** (1996). Demethylation-induced developmental pleiotropy in *Arabidopsis*. *Science* **273**: 654–657.
- Saito, Y., Tsuji, J., and Mituyama, T.** (2014). Bisulfighter: accurate detection of methylated cytosines and differentially methylated regions. *Nucleic Acids Res.* **42**: e45.
- Sasaki, E., Kawakatsu, T., Ecker, J.R., and Nordborg, M.** (2019). Common alleles of CMT2 and NRPE1 are major determinants of CHH methylation variation in *Arabidopsis thaliana*. *PLoS Genet.* **15**: e1008492.
- Saze, H., Mittelsten Scheid, O., and Paszkowski, J.** (2003). Maintenance of CpG methylation is essential for epigenetic inheritance during plant gametogenesis. *Nat. Genet.* **34**: 65–69.
- Schmitz, R.J., Schultz, M.D., Lewsey, M.G., O'Malley, R.C., Urich, M.A., Libiger, O., Schork, N.J., and Ecker, J.R.** (2011). Transgenerational epigenetic instability is a source of novel methylation variants. *Science* **334**: 369–373.
- Schmitz, R.J., Schultz, M.D., Urich, M.A., Nery, J.R., Pelizzola, M., Libiger, O., Alix, A., McCosh, R.B., Chen, H., Schork, N.J., and Ecker, J.R.** (2013). Patterns of population epigenomic diversity. *Nature* **495**: 193–198.
- Schultz, M.D. et al.** (2015). Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature* **523**: 212–216.
- Shen, X., De Jonge, J., Forsberg, S.K.G., Pettersson, M.E., Sheng, Z., Hennig, L., and Carlborg, Ö.** (2014). Natural CMT2 variation is associated with genome-wide methylation changes and temperature seasonality. *PLoS Genet.* **10**: e1004842.
- Slowikowski, K.** ggrepel: Repel overlapping text labels away from each other (Github).
- Srivastava, A., Karpievitch, Y.V., Eichten, S.R., Borevitz, J.O., and Lister, R.** (2019). HOME: a histogram based machine learning approach for effective identification of differentially methylated regions. *BMC Bioinformatics* **20**: 253.
- Stacklies, W., Redestig, H., Scholz, M., Walther, D., and Selbig, J.** (2007). pcaMethods - a bioconductor package providing PCA methods for incomplete data. *Bioinformatics* **23**: 1164–1167.

- Storey, J.D. and Tibshirani, R.** (2003). Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U. S. A.* **100**: 9440–9445.
- Stroud, H., Ding, B., Simon, S.A., Feng, S., Bellizzi, M., Pellegrini, M., Wang, G.-L., Meyers, B.C., and Jacobsen, S.E.** (2013a). Plants regenerated from tissue culture contain stable epigenome changes in rice. *eLife* **2**: e00354.
- Stroud, H., Do, T., Du, J., Zhong, X., Feng, S., Johnson, L., Patel, D.J., and Jacobsen, S.E.** (2014). Non-CG methylation patterns shape the epigenetic landscape in *Arabidopsis*. *Nat. Struct. Mol. Biol.* **21**: 64–72.
- Stroud, H., Greenberg, M.V.C., Feng, S., Bernatavichute, Y.V., and Jacobsen, S.E.** (2013b). Comprehensive analysis of silencing mutants reveals complex regulation of the *Arabidopsis* methylome. *Cell* **152**: 352–364.
- Sun, D., Xi, Y., Rodriguez, B., Park, H.J., Tong, P., Meong, M., Goodell, M.A., and Li, W.** (2014). MOABS: model based analysis of bisulfite sequencing data. *Genome Biol.* **15**: R38.
- Tedeschi, F., Rizzo, P., Huong, B.T.M., Czihal, A., Rutten, T., Altschmied, L., Scharfenberg, S., Grosse, I., Becker, C., Weigel, D., Bäumlein, H., and Kuhlmann, M.** (2019). EFFECTOR OF TRANSCRIPTION factors are novel plant-specific regulators associated with genomic DNA methylation in *Arabidopsis*. *New Phytol.* **221**: 261–278.
- Vaughn, M.W. et al.** (2007). Epigenetic natural variation in *Arabidopsis thaliana*. *PLoS Biol.* **5**: e174.
- Wassenegger, M., Heimes, S., and Sanger, H.L.** (1994). An infectious viroid RNA replicon evolved from an in vitro-generated non-infectious viroid deletion mutant via a complementary deletion in vivo. *EMBO J.* **13**: 6172–6177.
- Wibowo, A. et al.** (2018). Partial maintenance of organ-specific epigenetic marks during plant asexual reproduction leads to heritable phenotypic variation. *Proc. Natl. Acad. Sci. U. S. A.* **115**: E9145–E9152.
- Wibowo, A., Becker, C., Marconi, G., Durr, J., Price, J., Hagmann, J., Papareddy, R., Putra, H., Kageyama, J., Becker, J., Weigel, D., and Gutierrez-Marcos, J.** (2016). Hyperosmotic stress memory in *Arabidopsis* is mediated by distinct epigenetically labile sites in the genome and is restricted in the male germline by DNA glycosylase activity. *eLife* **5**: e13546.
- Wickham, H.** (2009). *ggplot2: Elegant Graphics for Data Analysis* (Springer, New York, NY).
- Wilkins, D.** *gggenes: Draw gene arrow maps in ggplot2* (Github).
- Woste, M., Leitao, E., Laurentino, S., Horsthemke, B., Rahmann, S., and Schroder, C.** (2020). wg-blimp: an end-to-end analysis pipeline for whole genome bisulfite sequencing data. *BMC Bioinformatics* **21**: 169.
- Wurmus, R., Uyar, B., Osberg, B., Franke, V., Gosdschan, A., Wreczycka, K., Ronen, J., and Akalin, A.** (2018). PiGx: reproducible genomics analysis pipelines with GNU Guix. *Gigascience* **7**.
- Zemach, A., McDaniel, I.E., Silva, P., and Zilberman, D.** (2010). Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science* **328**: 916–919.
- Zemach, A., Yvonne Kim, M., Hsieh, P.-H., Coleman-Derr, D., Eshed-Williams, L., Thao,**

bioRxiv preprint doi: <https://doi.org/10.1101/2022.01.06.475031>; this version posted January 6, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

K., Harmer, S.L., and Zilberman, D. (2013). The Arabidopsis nucleosome remodeler DDM1 allows DNA methyltransferases to access H1-containing heterochromatin. *Cell* **153**: 193–205.

Zhang, Y., Jang, H., Xiao, R., Kakoulidou, I., Piecyk, R.S., Johannes, F., and Schmitz, R.J. (2021). Heterochromatin is a quantitative trait associated with spontaneous epiallele formation. *Nat. Commun.* **12**: 1–13.

Zhong, S., Fei, Z., Chen, Y.-R., Zheng, Y., Huang, M., Vrebalov, J., McQuinn, R., Gapper, N., Liu, B., Xiang, J., Shao, Y., and Giovannoni, J.J. (2013). Single-base resolution methylomes of tomato fruit development reveal epigenome modifications associated with ripening. *Nat. Biotechnol.* **31**: 154–159.

Ziller, M.J., Gu, H., Müller, F., Donaghey, J., Tsai, L.T.-Y., Kohlbacher, O., De Jager, P.L., Rosen, E.D., Bennett, D.A., Bernstein, B.E., Gnirke, A., and Meissner, A. (2013). Charting a dynamic DNA methylation landscape of the human genome. *Nature* **500**: 477–481.

Supplemental Material to:

MethylScore, a pipeline for accurate and context-aware identification of differentially methylated regions from population-scale plant WGBS data

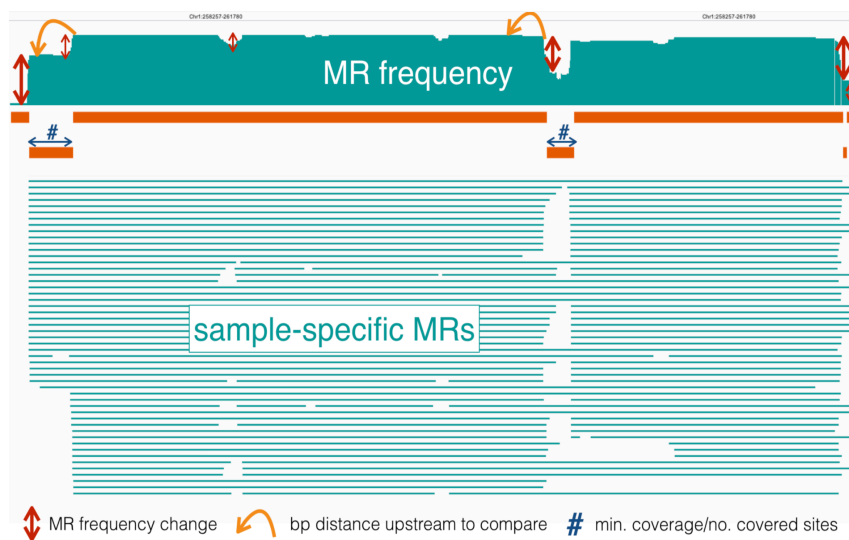
Patrick Hüther, Jörg Hagmann, Adam Nunn, Rahul Pisupati, Ioanna Kakoulidou, David Langenberger, Detlef Weigel, Frank Johannes, Sebastian J. Schultheiss, & Claude Becker

This supplemental material file contains:

- Supplement Figures 1-5

bioRxiv preprint doi: <https://doi.org/10.1101/2022.01.06.475031>; this version posted January 6, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

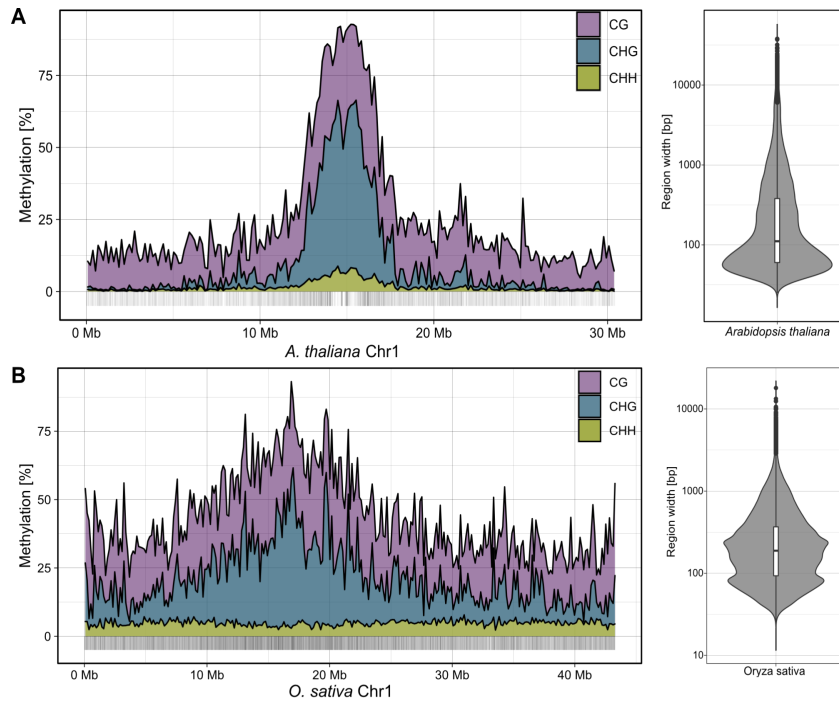
Supplemental Figure 1



Supplementary Figure 1: MR-based strategy to determine candidate DMRs.

The top panel shows a methylation region (MR) frequency barplot along the genome, summarizing the counts of all sample-specific MRs (the MRs are displayed as horizontal petrol-colored lines in the middle). Three parameters specify the borders of candidate differentially methylated regions (DMRs; orange horizontal bars): 1) the degree of MR frequency changes along the genome (marked by red vertical arrows), 2) the upstream distance to which MR frequency is compared to (marked by orange-colored bent arrows), and 3) a minimum number and coverage of cytosines (blue arrows indicating the length of candidate DMRs).

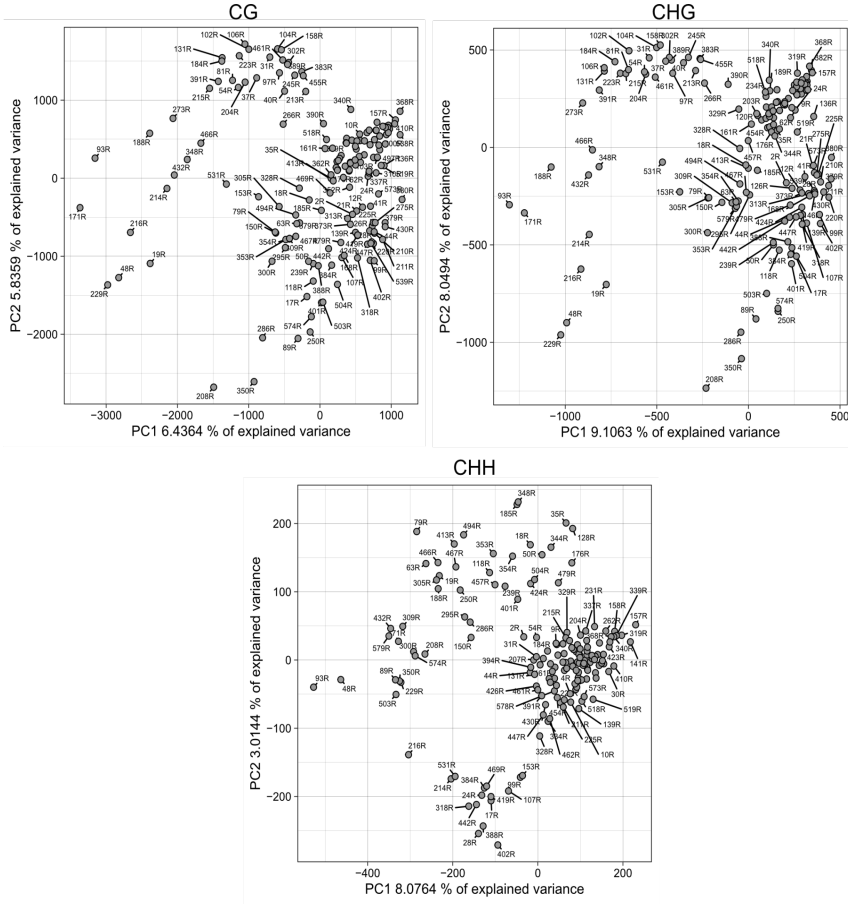
Supplemental Figure 2



Supplementary Figure 2: Genomic distribution of DNA methylation and methylated regions (MRs).

Lineplots (left) show the distribution of average DNA methylation in 150 kb bins along single chromosomes of *A. thaliana* (A; data from (Kawakatsu et al., 2016)) and *O. sativa* (B; data from (Stroud et al., 2013a)). MRs segmented by MethylScore are indicated as vertical bars along the x-axis, with corresponding length distributions shown as violin plots (right).

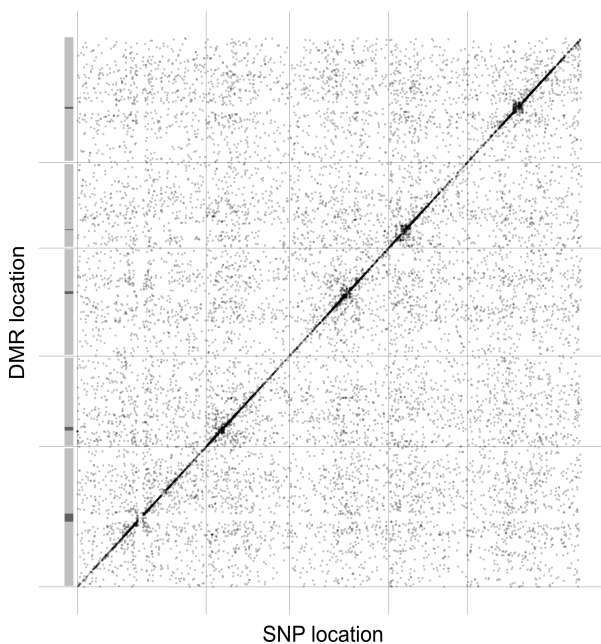
Supplemental Figure 4



Supplemental Figure 4: DMR identification across epigenetic recombinant inbred lines (epiRILs). Principal component analysis (PCA) of mean methylation rates within 46,323 CG-, 8,821 CHG- and 4,121 CHH-DMRs identified by MethylScore across a population of 169 epiRILs.

bioRxiv preprint doi: <https://doi.org/10.1101/2022.01.06.475031>; this version posted January 6, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplemental Figure 5



Supplementary Figure 5: GWA analysis on region-level methylation rate averages in CG.

3 Repression of CHROMOMETHYLASE 3 Prevents Epigenetic Collateral Damage in Arabidopsis

status	<i>published</i> eLife 10:e69396 (2021) doi:10.7554/eLife.69396
authors	Ranjith K. Papareddy (R.K.P.), Katalin Páldi (K.P.), Anna D. Smolka (A.D.S.), Patrick Hüther (P.H.), Claude Becker (C.B.) and Michael D. Nodine (M.D.N.)
contributions	R.K.P., Conceptualization, Software, Formal analysis, Investigation, Methodology, Writing - original draft, Writing - review and editing; K.P., Investigation, Methodology; A.D.S., Investigation; P.H., Software; C.B., Supervision; M.D.N., Conceptualization, Supervision, Funding acquisition, Writing - original draft, Writing - review and editing



RESEARCH ARTICLE



Repression of CHROMOMETHYLASE 3 prevents epigenetic collateral damage in *Arabidopsis*

Ranjith K Papareddy^{1†*}, Katalin Páldi^{1†}, Anna D Smolka¹, Patrick Hühner^{1,2}, Claude Becker^{1,2}, Michael D Nodine^{1,3*}

¹Gregor Mendel Institute (GMI), Austrian Academy of Sciences, Vienna Biocenter (VBC), Dr. Bohr-Gasse 3, Vienna, Austria; ²Genetics, LMU Biocenter, Ludwig-Maximilians University, Martinsried, Germany; ³Laboratory of Molecular Biology, Wageningen University, Wageningen, Netherlands

Abstract DNA methylation has evolved to silence mutagenic transposable elements (TEs) while typically avoiding the targeting of endogenous genes. Mechanisms that prevent DNA methyltransferases from ectopically methylating genes are expected to be of prime importance during periods of dynamic cell cycle activities including plant embryogenesis. However, virtually nothing is known regarding how DNA methyltransferase activities are precisely regulated during embryogenesis to prevent the induction of potentially deleterious and mitotically stable genic epimutations. Here, we report that microRNA-mediated repression of CHROMOMETHYLASE 3 (CMT3) and the chromatin features that CMT3 prefers help prevent ectopic methylation of thousands of genes during embryogenesis that can persist for weeks afterwards. Our results are also consistent with CMT3-induced ectopic methylation of promoters or bodies of genes undergoing transcriptional activation reducing their expression. Therefore, the repression of CMT3 prevents epigenetic collateral damage on endogenous genes. We also provide a model that may help reconcile conflicting viewpoints regarding the functions of gene-body methylation that occurs in nearly all flowering plants.

*For correspondence: rpapareddy1@g.ucla.edu (RKP); michael.nodine@wur.nl (MDN)

[†]These authors contributed equally to this work

Competing interests: The authors declare that no competing interests exist.

Funding: See page 18

Received: 13 April 2021

Accepted: 21 July 2021

Published: 23 July 2021

Reviewing editor: Richard Amasino, University of Wisconsin Madison, United States

© Copyright Papareddy et al. This article is distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use and redistribution provided that the original author and source are credited.

Introduction

Methylation of DNA encoding transposable elements (TE) is required to silence their expression and consequently prevent them from mobilizing and mutagenizing genomes (Kato et al., 2003; Law and Jacobsen, 2010). Complex mechanisms have evolved to balance the high degree of sensitivity needed to direct methylation and silencing of TEs with the precision required to prevent ectopic methylation of endogenous genes (Antunez-Sanchez et al., 2020; Ito et al., 2015; Papareddy et al., 2020; Saze and Kakutani, 2011; Williams et al., 2015; Zhang et al., 2020). However, little is known about the mechanisms of epigenome homeostasis during embryogenesis when organisms are particularly vulnerable to TE-induced mutagenesis, as well as the establishment of potentially deleterious epimutations that can persist through many cell divisions and even across generations (Henderson and Jacobsen, 2007; Mathieu et al., 2007; Probst et al., 2009; Saze et al., 2003; Mittelsten Scheid et al., 1998).

In *Arabidopsis thaliana* (Arabidopsis), most TEs are found in pericentromeric regions of the genome to which RNA polymerases have limited access (Arabidopsis Genome Initiative, 2000; Lippman et al., 2004; Zhang et al., 2006). These TE-enriched constitutive heterochromatic regions are characterized by high densities of cytosine methylation in symmetric (CG or CHG; H ≠ G) and asymmetric (CHH) contexts, as well as histone H3 lysine dimethylation (H3K9me2) and other transcriptionally repressive chromatin marks (Cokus et al., 2008; Lister et al., 2008; Stroud et al.,

2014). Symmetric DNA methylation and H3K9me2 also facilitate the stable propagation of silenced states through cell divisions (Jackson *et al.*, 2002; Lindroth *et al.*, 2001; Stroud *et al.*, 2013; Stroud *et al.*, 2014). METHYLTRANSFERASE 1 (MET1) maintains CG methylation through mitotic and meiotic cell divisions with high fidelity due to VARIANT IN METHYLATION 1/2/3 (VIM1/2/3) proteins that recognize hemi-methylated CG and recruit MET1 to methylate daughter strands (Feng *et al.*, 2010; Finnegan and Dennis, 1993; Ning *et al.*, 2020; Woo *et al.*, 2008). CG methylation can also recruit RNA Polymerase IV complexes required to produce 24-nt small interfering RNAs (siRNAs) that are then loaded onto Argonaute proteins and guide them to target loci by base-pairing with nucleic acids (Blevins *et al.*, 2015; Herr *et al.*, 2005; Papareddy *et al.*, 2020; Zhai *et al.*, 2015; Zilberman, 2003). This leads to the recruitment of DOMAINS REARRANGED METHYLTRANSFERASES 1/2 (DRM1/2) and results in de novo methylation of cytosines in all sequence contexts, including CHH, which is a hallmark feature of RNA-directed DNA methylation (RdDM) (Cao and Jacobsen, 2002; Stroud *et al.*, 2013; Wierzbicki *et al.*, 2008). However, RdDM is typically restricted from constitutive heterochromatin because it is inaccessible to DNA-dependent RNA polymerase IV and methyltransferases required for RdDM (Papareddy *et al.*, 2020; Zemach *et al.*, 2013). Instead, CHH methylation of constitutive heterochromatin is mediated by CHROMOMETHYLASE 2 (CMT2) that binds to H3K9me2 deposited by KRYPTONITE (KYP) and closely related SUPPRESSOR OF VARIATION 3-9 HOMOLOGUE PROTEIN 5/6 (SUVH5/6) methyltransferases (Stroud *et al.*, 2014; Zemach *et al.*, 2013). CHROMOMETHYLASE 3 (CMT3) also forms interlocking positive feedback loops with H3K9 methyltransferases (Du *et al.*, 2012; Jackson *et al.*, 2002; Lindroth *et al.*, 2001) but is more closely associated with the cell cycle and mediates CHG methylation (Ning *et al.*, 2020).

CMT3-mediated CHG methylation is largely deposited on TEs. However, CMT3 can also induce the ectopic methylation of protein-coding genes (Wendte *et al.*, 2019). Moreover, the introduction of Arabidopsis CMT3 transgenes into *Eutrema salsugineum*, which lost CMT3 millions of years ago (Bewick *et al.*, 2016), could occasionally reconstitute CG methylation on genes (Wendte *et al.*, 2019). The resulting gene-body methylation (gbM) could be stably maintained independent of the CMT3 transgene for several generations (Wendte *et al.*, 2019). However, it remains largely unknown how CMT3 is restricted to targeting heterochromatin, as well as the consequences of CMT3-induced hypermethylation of genes. Moreover, the functional significance of gbM in animals and plants has been intensely debated. Because methylated cytosines are mutagenic due to associated cytosine deamination (Shen *et al.*, 1992; Sved and Bird, 1990), features associated with gbM have been interpreted as evidence that gbM provides selective advantages that counterbalance this mutagenesis-imposed fitness penalty. For instance, gbM is depleted from transcription start and end sites (Tran *et al.*, 2005; Zhang *et al.*, 2006; Zilberman *et al.*, 2007), and it has recently been reported that gbM helps prevent transcription initiation from cryptic promoters located in gene bodies as initially proposed (Choi *et al.*, 2020; Zilberman *et al.*, 2007). Moreover, gbM tends to be enriched on constitutively expressed genes (Lister *et al.*, 2008; Niederhuth *et al.*, 2016; Takuno *et al.*, 2017; Zhang *et al.*, 2006), which would be consistent with gbM stabilizing gene expression by excluding certain histone variants (i.e. H2A.Z) from genes (Coleman-Derr and Zilberman, 2012) and generally enhancing gene expression (Muyle and Gaut, 2019; Shahzad *et al.*, 2021). Nevertheless, accumulating evidence is also consistent with gbM being a heritable by-product of CMT3-induced epimutations (Bewick *et al.*, 2016; Bewick *et al.*, 2019; Wendte *et al.*, 2019).

Consistent with the need to fine-tune the amount of CMT3 activities required to both silence TEs and prevent epimutations on genes, mechanisms exist that transcriptionally (Ning *et al.*, 2020) and post-translationally (Deng *et al.*, 2016) regulate CMT3, as well as remove H3K9me2 specifically from expressed genes (Inagaki *et al.*, 2010; Saze *et al.*, 2008). These and additional mechanisms are likely of utmost importance during embryo development when a proliferative morphogenesis phase produces the most fundamental cell lineages of the plant, including those that will eventually generate the gametes. Yet, how DNA methylation pathways are regulated during this phase of dynamic cell division to exquisitely balance the need for TE methylation with the prevention of potentially deleterious and stably inherited epimutations is virtually unknown.

Results

Cell division is linked with CG and CHG methylation through distinct mechanisms

MET1 and VIM1/2/3 are required for the faithful transmission of mCG across cell cycles (Feng et al., 2010; Finnegan and Dennis, 1993; Ning et al., 2020; Woo et al., 2008) and accordingly had increased transcript levels in rapidly dividing early embryos that also correlated well with transcripts encoding cell-cycle activators throughout embryogenesis (Figure 1A; Hofmann et al., 2019; Papareddy et al., 2020). More specifically, MET1 and VIM1/2/3 transcript levels peaked at the early heart stage and were reduced afterwards before plummeting at the mature green stage. These transcript developmental dynamics were also characteristic of transcripts encoding proteins involved in licensing DNA replication (e.g. Cyclins A2/B1, CDKB1-1, MINICHROMOSOME MAINTENANCE2), heterochromatin maintenance (e.g. DDM1) and DNA methylation (e.g. CMT3), but not randomized controls (Figure 1B, Figure 1—figure supplement 1A). Therefore, genes required for DNA methylation and heterochromatin maintenance are tightly correlated with cell-cycle activity during embryogenesis.

To test whether the patterns observed for transcripts regulating DNA methylation reflect DNA methylation dynamics, we computed differentially methylated cytosines (DMCs) across flowers, embryos and leaves (see Materials and methods). Similar to previous observations (Bouyer et al., 2017; Lin et al., 2017; Papareddy et al., 2020), 70% of DMCs occurred in the CHH context (Figure 1C). Consistent with dynamic expression patterns of MET1 and CMT3, substantial fractions of DMCs occurred in CG (20%) or CHG (10%) contexts, respectively. Therefore, DNA methylation is dynamically reconfigured in all sequence contexts during embryogenesis. In total, these symmetric DMCs represented 1185 CG (Supplementary file 1) and 1398 CHG (Supplementary file 2) differentially methylated regions (DMRs) covering 201 kb and 185.8 kb, respectively (Figure 1D, Supplementary file 1; see Materials and methods). Although a significant fraction of CG and CHG DMRs overlapped ($n = 183$; 7.1% of total), the vast majority of CG and CHG DMRs were located in non-overlapping genomic regions corresponding to euchromatic gene-rich and heterochromatic TE-rich regions of the genome, respectively (Figure 1D). Because CHG methylation can require CG methylation (Stroud et al., 2013), we tested whether the 15.1% of CHG DMRs overlapping CG DMRs require CG methylation. Leaves deficient in CG methylation did not have reduced CHG methylation in CHG DMRs regardless of whether or not they overlapped with CG DMRs (Figure 1E,F; data from Stroud et al., 2013). This indicates that CHG DMRs occur in distinct genomic regions and are largely independent of CG methylation.

Relative to floral bud samples, CG DMRs have slightly reduced methylation in preglobular embryos, followed by increased methylation until after the torpedo stage, when levels dramatically reduce in mature green embryos and recover in leaves (Figure 1G). By contrast, methylation levels of CHG DMRs are relatively stable between floral buds and early embryos, then decrease in late embryos, reaching a minimum in leaves (Figure 1H). Accordingly, changes in CG and CHG DMR methylation levels during development were significantly correlated with MET1 (Pearson's $R = 0.8$; p value = 0.03) and CMT3 (Pearson's $R = 0.74$; p value = 0.05) transcript levels, respectively (Figure 1—figure supplement 1B,C). Therefore, although cell division rates are correlated with symmetric DNA methylation dynamics, distinct mechanisms reconfigure CG or CHG methylation genome-wide during embryogenesis.

Genome-wide coordination of symmetric DNA methylation

Because DNA methylation is concentrated on TEs (Stroud et al., 2013; Zhang et al., 2006), we next investigated global developmental dynamics of TE methylation. CG methylation on both euchromatic and heterochromatic TEs was slightly reduced in preglobular embryos and then restored to the levels found in floral buds by the early heart stage (Figure 2A,B). Whereas CG methylation of euchromatic TEs was relatively constant for the remainder of embryogenesis, heterochromatic TEs had significantly increased methylation during late embryogenesis compared to post-embryonic tissues. Consistent with heterochromatin becoming highly condensed during embryo maturation (van Zanten et al., 2011), we found that CG hypermethylation in mature green compared to bent cotyledon embryos predominantly occurred in pericentromeric genomic regions rather than gene-

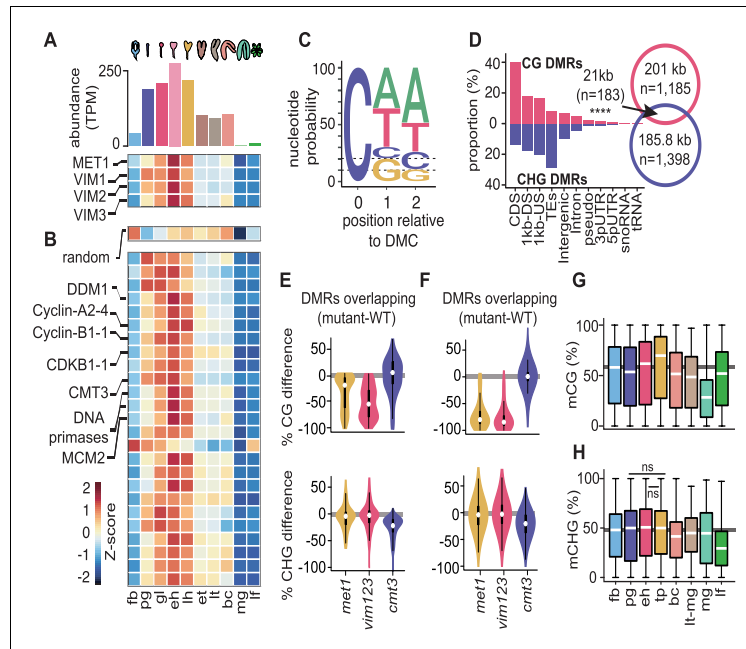


Figure 1. Cell division is linked with CG and CHG methylation through distinct mechanisms. (A) Bar chart depicting total abundance (top) and heat map of individual relative transcript levels (bottom) of genes involved in CG methylation in three biological replicates each of flowers, embryos, and leaves (Hofmann et al., 2019). fb, floral buds; pg, preglobular; gl, globular; eh, early heart; lh, late heart; et, early torpedoid; lt, late torpedoid; bc, bent cotyledon; mg, mature green; lf, leaf. (B) Heatmap showing developmental dynamics of permuted gene set (top) median values (i.e. 1000 iterations of random sampling of 25 genes) and top-25 genes co-varying with MET1, VIM1, VIM2, and VIM3 obtained by employing nearest neighbor algorithm calculated based on Euclidean distance between genes and centroid expression of MET1, VIM1, VIM2, and VIM3 (bottom). (C) Sequence logo representing nucleotide probability relative to differentially methylated cytosines (DMC). (D) Proportion of CG and CHG differentially methylated regions (DMRs) overlapping genomic features. Venn diagram showing overlap between CG and CHG DMRs. Significance overlap of DMRs determined by Fisher's Exact test p value < 0.0001 is indicated by ****. (E and F) Violin plot showing CG (top) and CHG (bottom) methylation differences between mutant and WT leaves for CHG DMRs overlapping (E) or not overlapping (F) with CG DMRs (Stroud et al., 2013). (G and H) Box plots of average weighted methylation of CG DMRs (n = 1185) and (G) CHG DMRs (n = 1398) during development. fb, floral buds; pg, preglobular; eh, early heart; tp, torpedoid (6 DAP) Pignatta et al., 2015; bc, bent cotyledon; lt-mg, late torpedoid-to-early mature green Hsieh et al., 2009; mg, mature green Bouyer et al., 2017; lf, leaf. fb, pg, eh, bc and lf were from Bouyer et al., 2017; Hsieh et al., 2009; Papareddy et al., 2020; Pignatta et al., 2015. Unless stated as not significant (ns), all combinations are significant with p values < 0.001 obtained by Mann-Whitney U test. Shaded horizontal line in the background represents the median methylation value of floral buds.

The online version of this article includes the following figure supplement(s) for figure 1:

Figure supplement 1. Characteristics of genes and differentially methylated regions co-expressed with symmetric methyltransferases.

rich chromosomal arms (Figure 2C). CG methylation was required for the production of 24-nt siRNAs from euchromatic TEs, but only marginally for heterochromatic TEs (Figure 2—figure supplement 1A, data from Lister et al., 2008). Conversely, the loss of 24-nt siRNAs in *nprpd1a* mutants only had negligible effects on CG methylation of both heterochromatic and euchromatic TEs (Figure 2—figure supplement 1B,C). Therefore, siRNA production from euchromatic regions of the genome requires CG methylation, but not vice versa.

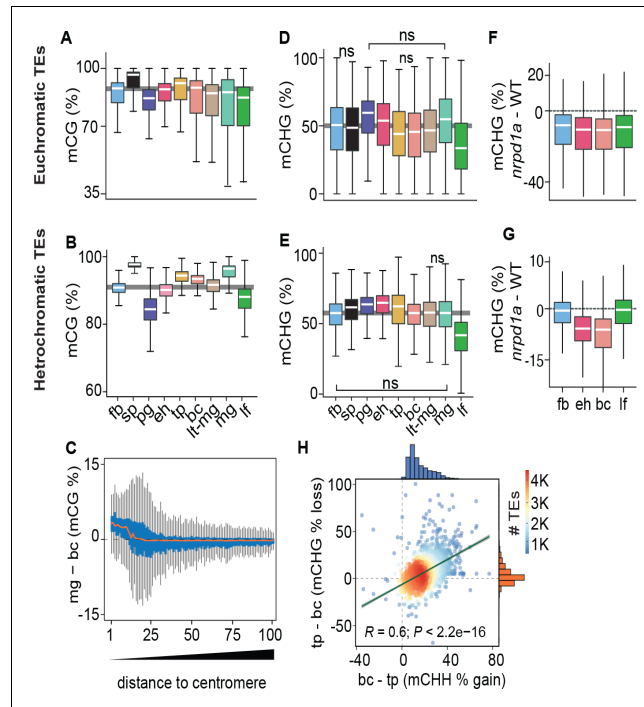


Figure 2. Genome-wide coordination of symmetric DNA methylation. (A and B) Boxplots of CG methylation percentages on euchromatic (A) and heterochromatic (B) TEs during development. fb, floral buds; sp, sperm (Ibarra et al., 2012); pg, preglobular; eh, early heart; tp, torpedo (6 DAP); bc, bent cotyledon; lt-mg, late torpedo-to-early mature green; mg, mature green; lf, leaf. Thick horizontal bars indicate medians, and the top and bottom edges of boxes represent the 75th and 25th percentiles, respectively. Shaded horizontal line in the background represents the median methylation value of floral buds. (C) Difference in CG methylation between mature green (mg) and bent cotyledon (bc) embryos were calculated in 1 kb genomic bins, which were divided into percentiles and sorted based on their distance to centromeres (1 and 100 being the tile closest and furthest from centromeres, respectively). Red color line indicates the median and the top and bottom edges of the blue colored boxes represent 75th and 25th percentiles, respectively. Vertical gray bars indicate 1.5X the interquartile range. (D and E) Boxplots of CHG methylation on euchromatic (D) and heterochromatic (E) TEs during development (key as in A). (F and G) Boxplots of CHG methylation differences between *nrpd1a* and WT (Col-0) tissues for euchromatic (F) and heterochromatic (G) TEs. (H) Scatterplot showing Pearson's correlation coefficients (R). Differences in mCHH and mCHG between bent cotyledon (bc) and torpedo stage (tp) embryos are shown on x- and y-axes, respectively. Histograms show the number of TEs in thousands (K). The online version of this article includes the following figure supplement(s) for figure 2:

Figure supplement 1. Relationships between MET1 and 24-nt siRNAs.

Global CHG methylation of euchromatic and heterochromatic TEs was higher in embryos compared to leaves (Figure 2D,E). Similar to previous observations for CHH methylation (Papareddy et al., 2020), siRNA-deficient *nrpd1a* mutant tissues had reduced CHG methylation on euchromatic or heterochromatic TEs in all or only embryonic samples, respectively (Figure 2F,G). Intriguingly, increased CHH methylation on heterochromatic TEs was significantly correlated with decreased CHG methylation during late stages of embryogenesis when cell division rates are reduced (Figure 2H). Therefore, CMT3-dependent CHG and CMT2-dependent CHH methylation of heterochromatic TEs are positively and negatively correlated with cell division rates, respectively.

Repression of CMT3 during embryogenesis regulates methylome dynamics

CMT3 is recruited to loci by binding to H3K9me2 deposited by SUVH4/5/6 histone methyltransferases (Du et al., 2012; Jackson et al., 2002; Lindroth et al., 2001; Stroud et al., 2014). CMT3 and KYP, which is the major SUVH4 H3K9 methyltransferase, were dynamically expressed according to patterns characteristic of other cell-cycle regulated genes and CHG methylation dynamics (Figure 1B,H, Figure 3A). More specifically, CMT3 and KYP were highly expressed in rapidly dividing early embryos and had reduced expression in late embryos until the mature stage, where they were barely detectable. Altogether, our results are consistent with the idea that the more rapid cell divisions in early embryos demand higher levels of CMT3 and KYP to maintain mCHG through the cell cycle. Moreover, IBM1, which encodes an H3K9me2 demethylase and prevents CMT3 recruitment to gene bodies (Miura et al., 2009; Saze et al., 2008), is dynamically expressed during embryogenesis in a pattern that strongly resembles CMT3 and KYP (Figure 3A). Therefore, co-expression of IBM1 with CMT3 and KYP likely helps limit ectopic H3K9me2 and methylated CHG on gene bodies during

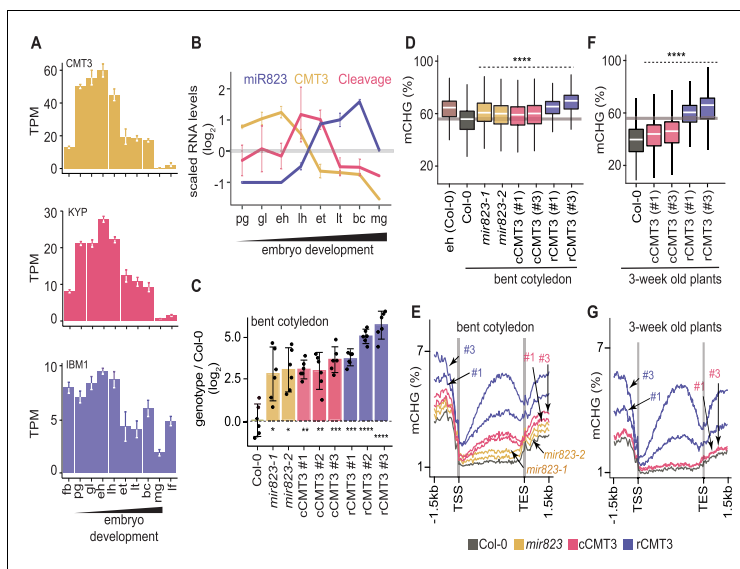


Figure 3. Repression of CMT3 during embryogenesis regulates methylome dynamics. (A) Barplots illustrating transcript levels of CMT3 (top), KYP (middle), and IBM1 (bottom) in flowers, embryos, and leaves. fb, floral buds; pg, preglobular; gl, globular; eh, early heart; lh, late heart; et, early torpedo; lt, late torpedo; bc, bent cotyledon; mg, mature green; lf, leaf. (B) Line graphs showing the relative RNA abundance of miR823 (blue), CMT3 RNA (yellow), and miR823:CMT3 cleavage products (pink). (C) Log₂-transformed relative CMT3 target transcript levels in bent cotyledon embryos (8 DAP; day after pollination) from WT plants (Col-0), or *cmt3-11* plants expressing either miR823-cleavable CMT3 (cCMT3) or miR823-resistant CMT3 (rCMT3) versions. Each dot represents the mean of two technical replicates of embryos and bars represent mean values. Error bars in A-C represent standard errors of the means of three biological replicates. Asterisks indicate whether the transcript levels observed in *mir823* mutant, cCMT3 and rCMT3 embryos were significantly different compared to WT (Two-tailed Student's t tests; ****, ***, **, and * represent p values < 0.0001, < 0.001, < 0.01, and < 0.05, respectively). Color-coded according to the key. (D) Boxplots of CHG methylation on transposons with ≥ 5 informative cytosines covered by ≥ 4 reads and classified as either euchromatic or heterochromatic in Papareddy et al., 2020. p-values < 0.0001 based on Mann-Whitney U tests of methylation differences between WT and either mutant or transgenic bent cotyledon embryos are represented by ****. (E) Metaplots of average CHG methylation percentages across genes bodies from transcription start sites (TSS) to transcription end sites (TES), 1.5 kb upstream and 1.5 kb downstream of genes in bent cotyledon embryos. Color-coded according to the key. (F and G) Boxplots of CHG methylation on transposons (F) and metaplots of CHG methylation on genes (G) in three-week old plants as described in D and E, respectively.

The online version of this article includes the following figure supplement(s) for figure 3:

Figure supplement 1. *mir823* mutants and effects of miR823-directed repression of CMT3.

embryogenesis as has been demonstrated during post-embryonic development (Inagaki *et al.*, 2017).

We previously found that miR823-directed cleavage of CMT3 transcripts is highly enriched in embryos directly after morphogenesis (Plotnikova *et al.*, 2019). In contrast to CMT3 transcript dynamics, miR823 accumulates during embryogenesis, and miR823:CMT3 cleavage products were enriched and significantly detected specifically at late heart and early torpedo stages precisely when CMT3 transcript levels were sharply decreasing (Figure 3B). Based on these observations, we hypothesized that miR823-mediated repression of CMT3 contributes to the reduced CHG methylation levels observed during late embryogenesis.

To test if miR823-directed repression of CMT3 transcripts reduces CHG methylation levels during embryogenesis, we generated deletions in the region of the *MIR823* locus encoding the mature miRNA (Figure 3—figure supplement 1A) and examined CMT3 transcript and CHG methylation levels. Both independently generated *mir823-1* and *mir823-2* mutants were confirmed as nulls (Figure 3—figure supplement 1B) and had significantly increased CMT3 levels relative to wild type (WT) in embryos at the bent cotyledon stage when CMT3 levels are normally reduced (Figure 3C). Consistent with miR823-directed cleavage of CMT3 being highly enriched in embryos, we did not observe increased CMT3 transcripts in either leaves or floral buds of *mir823* mutants (Figure 3—figure supplement 1D). Moreover, CHG, but not CG or CHH, methylation was increased on TEs in bent cotyledon embryos of both *mir823-1* and *mir823-2* mutants relative to WT (Figure 3D, Figure 3—figure supplement 1E).

As an independent approach, we used site-directed mutagenesis to introduce synonymous mutations in the miR823 target site within CMT3 transgene constructs that included 1.41 kb upstream and 0.73 kb downstream intergenic regions, and associated cis-regulatory elements (Figure 3—figure supplement 1C; see Materials and methods). As controls, we also generated CMT3 constructs without mutations, and introduced these miR823-cleavable CMT3 (cCMT3), as well as the miR823-resistant (rCMT3), constructs into *cmt3-11* mutant plants (Henderson and Jacobsen, 2008). CMT3 transcript levels were increased in rCMT3 relative to cCMT3 lines at the bent cotyledon stage (Figure 3C), but not in leaves or floral buds (Figure 3—figure supplement 1D), which further indicates that miR823-directed cleavage and repression of CMT3 is highly enriched in embryos transitioning between morphogenesis and maturation. CMT3 levels were also increased in cCMT3 and rCMT3 lines compared to Col-0 in embryos, leaves, and floral buds (Figure 3C, Figure 3—figure supplement 1D) suggesting that miR823 is not sufficient to repress transgenic CMT3 to the same extent as endogenous CMT3 transcripts. Although we cannot rule out that this is due to missing cis-regulatory repressive elements in the transgenes, increased gene dosage and positional effects of the transgenes seems more likely. Upstream and downstream intergenic regions were included in the CMT3 constructs (Figure 3—figure supplement 1G). Moreover, although relative transgene copy numbers were not significantly different across the independently generated cCMT3 and rCMT3 transgenic lines, they were higher than endogenous CMT3 in WT (Figure 4—figure supplement 1G,H). Nevertheless, it is clear that CMT3 levels are finely tuned during embryogenesis. Together with the analysis of *mir823* mutants and miR823-mediated CMT3 transcript cleavage products (Plotnikova *et al.*, 2019), these results strongly indicate that miR823 cleaves and represses CMT3 levels during mid-embryogenesis. Consistent with what we observed in *mir823* mutants, increased CMT3 transcript levels in cCMT3 and rCMT3 embryos resulted in CHG hypermethylation of TEs (Figure 3D) but did not globally influence CG or CHH methylation (Figure 4—figure supplement 1E,F). Remarkably, increased CMT3 transcript levels in *mir823* mutants, cCMT3 and most strikingly rCMT3 embryos were associated with ectopic CHG methylation on protein-coding gene bodies and flanking regions in bent cotyledon embryos (Figure 3E). Therefore, both TEs and genes are hypermethylated when CMT3 levels are not properly downregulated upon the morphogenesis-to-maturation transition during embryogenesis.

To test whether miR823-directed repression of CMT3 and prevention of CHG methylation of genes that we observed in embryos persists after embryogenesis, we next profiled methylomes of cCMT3 and rCMT3 plants 3 weeks after germination. We chose to focus on rCMT3 plants because of the large amount of hypermethylation observed in these lines during embryogenesis, and used cCMT3 plants as controls. Although TEs had increased CHG methylation levels in both cCMT3 lines relative to WT, protein-coding genes were not affected (Figure 3F,G). In stark contrast, TEs and genes were hypermethylated in both rCMT3 lines compared to cCMT3 or WT plants, and only

slightly reduced relative to the levels observed in rCMT3 bent cotyledon embryos (Figure 3F,G). Together with miR823-independent processes (e.g. IBM1 removal of H3K9me2), miR823-directed repression of CMT3 is therefore required to prevent the hypermethylation of protein-coding genes that can be maintained weeks after the completion of embryogenesis.

Chromatin features associated with CMT3-induced gene methylation

To yield insights into how genes are hypermethylated upon the derepression of CMT3, we determined whether certain genomic features were associated with CMT3-induced genic methylation. Toward this end, we first selected 22,637 nuclear-encoded protein-coding genes that had ≥ 5 methylC-seq reads overlapping CHG sites in rCMT3 line #3 and that were expressed (i.e. ≥ 1 TPM in any tissue based on Hofmann et al., 2019). We chose rCMT3 line #3 because it had the strongest genome-wide CHG hypermethylation and focussed on expressed genes to exclude those that may have TE-like features, which could confound analysis. We then used k-means clustering of the differences between rCMT3 line #3 and WT bent cotyledon embryos to partition this set of genes into four clusters (Figure 4—figure supplement 1A). These clusters were comprised of 1439–7882 genes (6.4–34.8% of total) and ranged from groups of genes that had no methylation changes (cluster 1) to those that were strongly hypermethylated with 3' biases (cluster 4) in rCMT3 compared to WT embryos (Figure 4A,B, Figure 4—figure supplement 1B,C). The same patterns were observed across these clusters in embryos from an independently generated rCMT3 transgenic (line #1), which indicates that CMT3-induced hypermethylation is not stochastic (Figure 4—figure supplement 1B, C).

TE-like methylated (teM) genes generally have non-CG methylation on their gene bodies without strong 3' biases (Kawakatsu et al., 2016; Bewick et al., 2016). To check whether rCMT3-induced genic CHG methylation is affected by teMs, we intersected our gene clusters with previously defined teMs (Bewick et al., 2016) and found that only 272 of 22,637 (0.012%) expressed genes overlapped teMs (Figure 4—figure supplement 1D). Hypermethylated clusters (clusters 3 and 4) contained more teMs compared to unmethylated or lowly methylated gene clusters (Figure 4—figure supplement 1E). However, rCMT3 embryos still had genic CHG hypermethylation and 3' biases after excluding teM genes, whereas, WT embryos remained devoid of CHG methylation (Figure 4—figure supplement 1F,G). Therefore, our analysis is not confounded by either TEs or teM genes. As expected, CMT3-induced hypermethylation predominantly occurred in the CMT3-preferred CWG context (Gouil and Baulcombe, 2016; Li et al., 2018) although hypermethylation was also found in CCG and slightly, but significantly in non-CHG contexts, including CG characteristic of gbM similar to previous observations (Figure 4—figure supplement 1H; Wendte et al., 2019).

Consistent with methyltransferases preferring nucleosome-rich DNA as substrates (Chodavarapu et al., 2010; Du et al., 2012), CMT3-induced hypermethylation was proportional to patterns of nucleosome occupancy and biased towards the 3' ends of gene bodies, which was highly similar to CG methylation (Figure 4C, Figure 4—figure supplement 1I). Nucleosome spacing is promoted by linker histone 1 (H1) (Choi et al., 2020; Fan et al., 2003) and CMT3-induced CHG hypermethylation was proportional to H1 levels across gene bodies (Figure 4—figure supplement 1J). Because nucleosome occupancy was not as readily distinguishable between clusters of affected genes (i.e. clusters 2–4) (Figure 4—figure supplement 1I,J), we hypothesized that histone variants conferring differential nucleosome stabilities and chromatin accessibility may influence ectopic CMT3-induced hypermethylation (Osakabe et al., 2018). Indeed, CHG hypermethylation across the four groups was positively correlated with levels of the stable histone variants H2A, H2A.X and most notably H2A.W that was recently shown to be required for CHG methylation (Figure 4E,F, Figure 4—figure supplement 1K; Bourguet et al., 2021; Yelagandula et al., 2014). CMT3-induced CHG hypermethylation was also tightly associated with transcriptionally repressive H3K9me2 marks, which are required for interdependent feedback loops with CMT3 (Figure 4G). It was inversely related to H2A.Z (Figure 4H) and marks indicative of active transcription including H3K4me3 and H3K9ac (Figure 4I, Figure 4—figure supplement 1L). Further suggesting that deregulated CMT3 prefers features typically associated with inaccessible chromatin, genes with CHG hypermethylation had reduced chromatin accessibility (Figure 4—figure supplement 1M) and were generally closer to heterochromatic centromeres (Figure 4—figure supplement 1N). Moreover, the most CMT3-induced hypermethylated genes (i.e. cluster 4) were also substantially hypermethylated in *ddm1* mutants (Figure 4J) that have increased heterochromatic accessibility (Figure 4K) and decreased

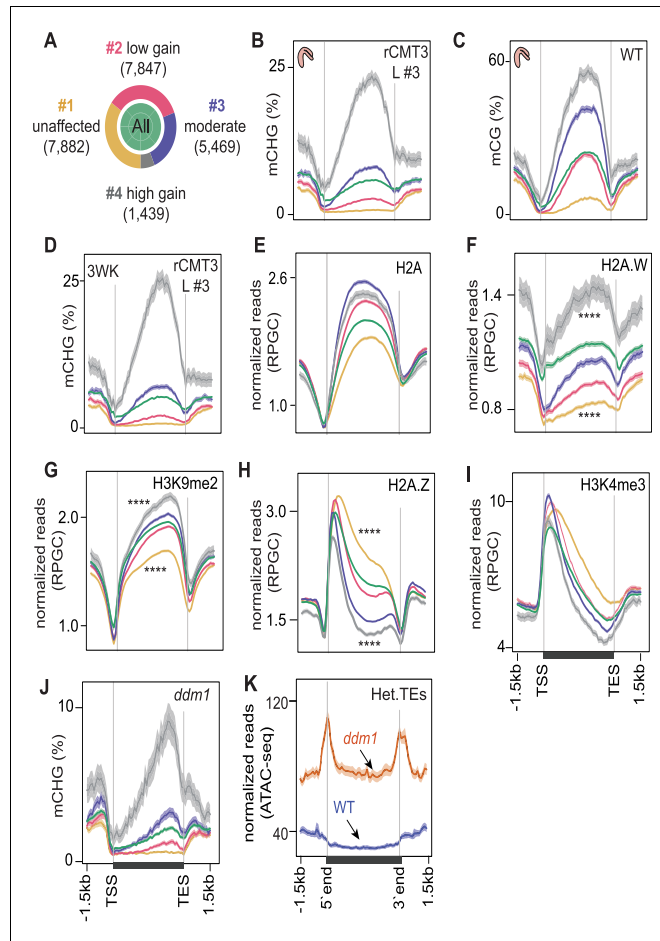


Figure 4. Chromatin features associated with CMT3-induced gene methylation. (A) Proportion of genes in each cluster partitioned using k-means clustering algorithm based on differences in mCHG between rCMT3 (line #3) and WT embryos. Unaffected genes (yellow), low mCHG gain genes (red), moderate mCHG gain genes (blue), and high mCHG gain genes (gray). Green inner circle represents all expressed genes. (B–D) Metaplots showing mCHG on gene clusters in bent cotyledon embryos from rCMT3 line #3 (L #3) (B), mCG on gene clusters in WT bent cotyledon embryos (C) and mCHG in rCMT3 line #3 3-week-old plants (3WK) (D). Shaded ribbons in metaplots represent standard deviations. (E–I) Metaplots showing normalized reads per genomic content (RPGC) average values of histone variant H2A (E), H2A.W (F) (Yelagandula et al., 2014), H3K9me2 (G) (Stroud et al., 2014), H2A.Z (H) (Yelagandula et al., 2014), and H3K4me2 (I) (Maher, 2020). p Values < 0.0001 obtained by Mann-Whitney U test based on differences between genes in cluster 1 or four compared to all genes is represented by ****. (J) Metaplots showing mCHG on gene clusters in seventh generation *ddm1* mutants (Stroud et al., 2013). (K) Normalized ATAC-seq reads (Zhong et al., 2021) representing accessibility of heterochromatic TEs (Het.TEs) in WT and *ddm1* mutants as defined in Papareddy et al., 2020.

The online version of this article includes the following figure supplement(s) for figure 4:

Figure supplement 1. Partitioning of CMT3-induced hypermethylated genes and associated chromatin features.

stability (Mathieu et al., 2003; Soppe et al., 2002; Zhong et al., 2021). Although CMT3-induced CHG hypermethylation was strongly associated with CG gene-body methylation (gbM), both the independence of developmental mCHG DMRs (Figure 1E,F) and the gain of mCHG being associated with proportional loss of mCG over genes in *ddm1* mutants (Figure 4—figure supplement 1O; Ito et al., 2015; Stroud et al., 2013; Zemach et al., 2013) indicate that mCG is not strictly required for ectopic CHG hypermethylation of genes. Instead, the associations between chromatin features of genes and their propensity for CMT3-induced hypermethylation altogether suggest that excessive CMT3 is ectopically recruited to genic chromatin characterized by nucleosome stability and inaccessibility.

Impact of CMT3-induced hypermethylation on gene expression

Because CHG methylation of TEs contributes to their repression (Stroud et al., 2014), we tested whether CMT3-induced ectopic CHG hypermethylation of protein-coding genes also represses their expression levels. Namely, we performed mRNA-seq on three biological replicates of WT and rCMT3 (line #s 1 and 3) bent cotyledon embryos. Principal component analysis revealed that WT and rCMT3 biological replicates clustered according to genotype and in similar positions along the dominant principal component axis corresponding to developmental time (Figure 5A). This indicates that our mRNA-seq datasets captured gene expression variation inherent to WT and rCMT3 genotypes, as well as that our staging was accurate. Differences in global transcript levels were not observed across the four clusters with increasing levels of CMT3-induced CHG methylation suggesting that ectopic CHG methylation alone was not sufficient to globally repress gene expression (Figure 5B, Figure 5—figure supplement 1A). We then identified 916 genes that were differentially expressed between rCMT3 and WT embryos (i.e. ≥ 2 -fold differences and adj. p values ≤ 0.01 ; see Materials and methods) (Figure 5—figure supplement 1B,C and Supplementary file 3). Differentially expressed genes (DEGs), defined by comparing either rCMT3 line #1 or rCMT3 line #3 with WT, were commonly detected in both independently generated lines with 87.5% of genes overlapping (Figure 5—figure supplement 1C). In both rCMT3 lines, DEGs were less hypermethylated compared to all expressed genes, which indicates that the vast majority of changes in gene expression observed upon up-regulation of CMT3 were not directly due to their hypermethylation (Figure 5—figure supplement 1D). We then examined whether hypermethylation affects a subset of genes by computing DMRs in rCMT3 compared to WT bent cotyledon embryos and identified 4603 (97% of total) and 127 (3% of total) CHG hypermethylated and hypomethylated DMRs, respectively (Supplementary file 4; see Materials and methods). Further suggesting that CHG hypermethylation has minimal direct consequences on the expression of most genes under the conditions examined, we found that only a small but significant number of the down-regulated genes (including 1.5 kb regions flanking their transcriptional units) overlapped DMRs (21 of 542, 3.8% of total; Fisher's exact test, p value = 1.29×10^{-5}) (Figure 5—figure supplement 1E). Consistent with CMT3-induced hypermethylation repressing their expression, the DMRs overlapping these 21 down-regulated genes were significantly CHG hypermethylated compared to genomic bins (Figure 5C). Moreover, the stronger-expressing rCMT3 line #3 had significantly higher CHG methylation compared to rCMT3 line #1 (Figure 5C). This further supports that increased CMT3 levels lead to more ectopic CHG methylation (Inagaki et al., 2010; Inagaki et al., 2017). However, the transcript levels of these 21 genes were only moderately reduced in rCMT3 line #3 compared to rCMT3 line #1, suggesting non-linear relationships between gene hypermethylation and transcript levels (Figure 5—figure supplement 1F). Strikingly, transcripts corresponding to these 21 CMT3-induced hypermethylated and down-regulated genes were rapidly increasing when embryos were transitioning to the maturation phase (Figure 5D). Moreover, 10 of these 21 genes (Fisher's exact test, P value = 1.49×10^{-13}), were among a group of 381 genes previously identified to also be rapidly activated at these time points (Figure 5—figure supplement 1G; Hofmann et al., 2019). Nearly half of these 381 genes (n = 183, 48%) were also among the 563 significantly down-regulated genes in rCMT3 embryos compared to WT. Altogether, our expression and methylation analyses suggest that when CMT3 is not properly repressed it can induce ectopic hypermethylation of genes. Furthermore, we suggest that CMT3-induced hypermethylation of gene promoters or bodies can reduce the steady state levels of transcripts from genes that are in the process of switching from silent to active transcriptional states. However, additional experiments are required to directly test whether gene-body CHG methylation can repress gene expression.

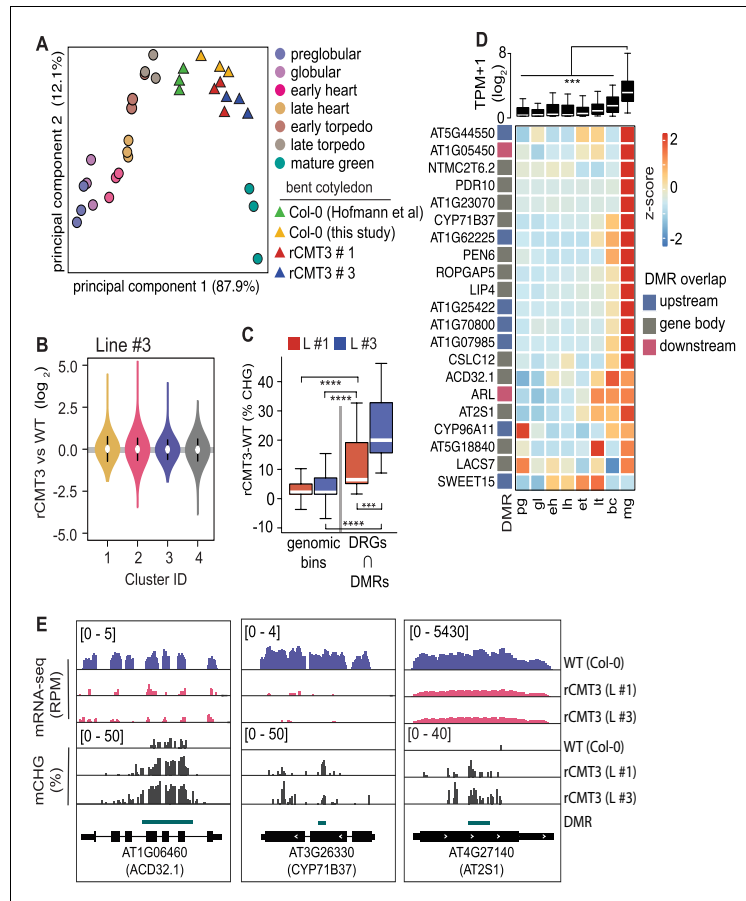


Figure 5. Impact of CMT3-induced hypermethylation on gene expression. (A) Principal component analysis of mRNA-seq from three biological replicates of rCMT3 and WT (Col-0) bent cotyledon embryos generated in this study along with floral buds, embryos, leaves, from Hofmann et al., 2019 and color-coded according to the key. (B) Violin plot showing transcript fold changes in rCMT3 (line #3) compared to WT (Col-0) bent cotyledon embryos per cluster as defined in Figure 4A. (C) Boxplot showing difference in methylation comparing rCMT3 to WT in down-regulated genes (DRG) intersecting with DMRs and similarly sized genomic bins of 213 bp as controls. p Values < 0.001 and < 0.0001 based on Mann-Whitney U tests are represented by *** and ****, respectively. (D) Boxplot (top) and heatmap (bottom) of transcript levels of DRGs intersecting DMRs during embryogenesis. p Values < 0.001 based on differences in transcript levels between mature green (mg) and all other stages of embryogenesis based on Mann-Whitney U test are represented by ***. (E) Integrative genome viewer (IGV) screenshot of representative downregulated genes associated with DMRs.

The online version of this article includes the following figure supplement(s) for figure 5:

Figure supplement 1. Additional information regarding the influence of CMT3-induced hypermethylation on gene expression.

Discussion

DNA methylation is faithfully propagated across cell cycles by methyltransferases to ensure robust silencing of TEs (Borges et al., 2021; Law and Jacobsen, 2010; Mathieu et al., 2007; Ning et al., 2020; Probst et al., 2009; Saze et al., 2003). However, it is not well understood how DNA

methyltransferases are regulated following periods of rapid division to prevent off-targeting of genes and their consequential repression. Cell division rates are highly dynamic during *Arabidopsis* embryogenesis. We found that the expression of MET1 and CMT3 methyltransferases and corresponding CG and CHG methylation are intricately linked to mitotic indices through distinct mechanisms (Figure 1). Moreover, miR823-mediated cleavage and repression of CMT3 following the proliferative early phase of embryogenesis helps prevent excess CMT3 from ectopically methylating protein-coding genes that can persist for weeks afterwards (Figure 3). CMT3-induced hypermethylation of genes was highly associated with features conferring nucleosome stability (Figure 4) and resulted in the repression of genes that are transcriptionally activated (Figure 5). Repression of CMT3 following a period when it is needed in high quantity to keep pace with TE methylation therefore prevents CMT3 from ectopically targeting protein-coding genes for methylation. This resulting epigenetic collateral damage on protein-coding genes appears to negatively affect gene expression. Our results are consistent with the model that CMT3-induced epimutations give rise to CG gene-body methylation (gbM) that can be maintained by MET1 across many generations (Wendte et al., 2019).

Complex mechanisms are required to specifically silence mutagenic TEs rather than endogenous genes (Antunez-Sanchez et al., 2020; Deng et al., 2016; Lee et al., 2021; Lister et al., 2008; Papareddy et al., 2020; Saze and Kakutani, 2011; Williams et al., 2015; Zhang et al., 2020). Mechanisms regulating epigenome homeostasis are of paramount importance during *Arabidopsis* embryogenesis due to highly dynamic cell cycle and transcriptional activities, as well as the establishment of cell lineages that will produce all future cell types including the gametes. MET1 and CMT3 methyltransferases are required for TE methylation (Kato et al., 2003; Stroud et al., 2014) and are expressed at high levels during early embryogenesis likely because this is a period of rapid cell division. CHG and CHH methylation exhibit opposite developmental dynamics depending on the tissue's mitotic index (Figure 1, Figure 2, Figure 6; Papareddy et al., 2020). When embryos are transitioning to stages with reduced cell division, decreased CMT3-mediated CHG methylation is correlated with increased CMT2-mediated CHH methylation (Figure 2H). Unlike CMT2, CMT3 can also target protein-coding genes for CHG methylation (Stroud et al., 2014) and lead to the recruitment of transcriptionally repressive H3K9me2 methyltransferases such as KYP (Du et al., 2014; Jackson et al., 2002; Lindroth et al., 2001). Therefore, handing over TE silencing to CMT2-dependent CHH methylation in cells with reduced division rates likely reduces ectopic methylation of protein-coding genes. In addition to what we observed during embryogenesis, varying degrees of mitotic indices across development can readily explain the genome-wide patterns of non-CG methylation reported thus far (Borges et al., 2021; Calarco et al., 2012; Gutzat et al., 2020; Ji et al., 2019; Kawakatsu et al., 2016, Kawakatsu et al., 2017; Lin et al., 2017; Narsai et al., 2017; Papareddy and Nodine, 2021).

CMT3, KYP and their corresponding DNA and histone methylation marks form interdependent feedback loops that perpetuate silencing through cell divisions (Du et al., 2015; Ning et al., 2020). Consistent with the transcription-coupled H3K9me2 demethylase IBM1 breaking these loops and preventing ectopic CHG hypermethylation of genes, we found that CMT3, KYP, and IBM1 were highly expressed during early embryogenesis (Figure 3). After this rapidly dividing morphogenesis phase, transcripts from CMT3, KYP, and IBM1 decrease, and miR823 directs the cleavage and repression of excess CMT3 to help prevent hypermethylation of protein-coding genes (Figure 3). Excess CMT3 induces CHG methylation on distinct regions of protein-coding genes that are characteristic of stable nucleosomes including transcriptionally repressive H3K9me2 marks that bind to CMT3. Although the distribution of CMT3-induced CHG hypermethylation is strikingly similar to CG gene-body methylation of genes (Figure 4B,C), this appears to be due to common targeting mechanisms by CMT3 and MET1 rather than a strict prerequisite of CG. In fact, mutants with reduced CG methylation (Figure 4—figure supplement 10; Jacobsen and Meyerowitz, 1997; Lister et al., 2008; Saze and Kakutani, 2007; Stroud et al., 2013) or species largely devoid of genic CG methylation (Wendte et al., 2019) can still recruit CHG on genes. CMT3-induced CHG methylation of genes that we observed in rCMT3 transgenic plants was similar to ectopic gain of genic mCHG in *ddm1* mutants (Figure 4). Notably, heterochromatin becomes destabilized in *ddm1* mutants (Figure 4K; Mathieu et al., 2003; Soppe et al., 2002) and CMT3 prefers features associated with stable (Figure 4E,F, Figure 4—figure supplement 1K; Bourguet et al., 2021; Osakabe et al., 2018; Yelagandula et al., 2014) over unstable nucleosomes such as H2A.Z with active marks

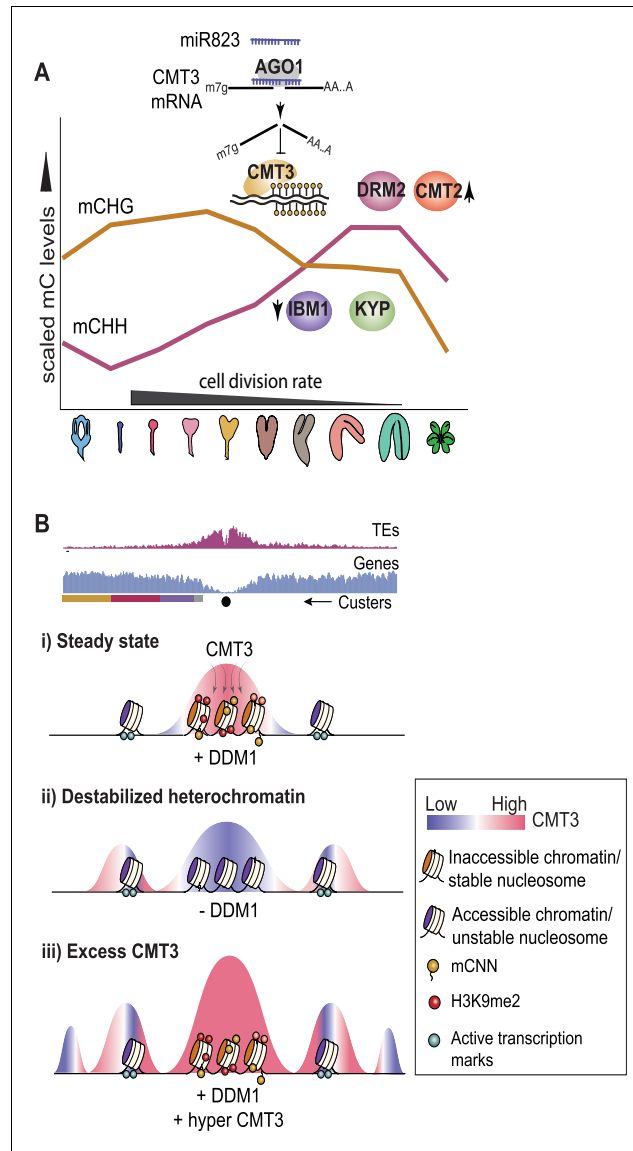


Figure 6. Models for CMT3 regulation during periods of fluctuating cell division rates and destabilized chromatin. **(A)** Model of non-CG methylation dynamics during embryo development and corresponding regulatory mechanisms. **(B)** Model for how CMT3 equilibrium is maintained to restrict its activity to heterochromatin. Density of transposable elements (TEs) (top; red) and genes (middle; blue) on chromosome 1. Cartoon illustration of gene cluster location (bottom) according to key in Figure 4A. Black dot represents the centromere. (i) In steady state, stable nucleosomes along with H3K9me2 and DNA methylation provides positive reinforcement to sequester CMT3 to constitutive heterochromatin. (ii) Loss of DDM1 results in destabilized and accessible heterochromatin (Zhong et al., 2021), characterized by loss of H3K9me2 and stable nucleosomes (Osakabe et al., 2021). Accessible chromatin or DNA without stable nucleosomes is no longer a preferable substrate for CMT3 and results in CHG hypomethylation of TEs. CMT3 will now be readily available and redirected to genic regions where it

Figure 6 continued on next page

Figure 6 continued

induces ectopic CHG methylation in proportion to the levels of stable nucleosomes and chromatin marks. (iii) Excess levels of CMT3 causes genome-wide CHG hypermethylation with a preference for stable nucleosomes associated with repressive marks that tend to be in regions closer to centromeres compared to chromosomal arms.

(Figure 4H,I, Figure 4—figure supplement 1L). Although destabilization of heterochromatin has been inversely correlated with genic CHG methylation (Ito et al., 2015; Zhang et al., 2020), chromatin features underlying this mechanism are unclear. Therefore, we propose that destabilization of heterochromatin in *ddm1* mutants redirects CMT3 to genic regions with stable nucleosomes. Therefore, factors such as DDM1 that stabilize heterochromatin may be yet another mechanism required to regulate CMT3 activity in order to achieve proper epigenome homeostasis (Figure 6).

CMT3-induced CHG hypermethylation of genes did not globally affect steady state transcript levels (Figure 5). However, we observed exceptional association between CHG hypermethylation and repression of genes that switch from transcriptionally inactive to active states. Because IBM1-mediated removal of H3K9me2 marks is coupled to transcription (Inagaki et al., 2017), it is possible that CMT3-induced methylation can form feedback loops with H3K9me2 methyltransferases when genes are transcriptionally inert. However, when genes are switched on, H3K9me2 could repress initial rounds of transcription before it is removed by IBM1. Accordingly, it may be difficult to detect the effects of ectopic CHG methylation on gene expression when quantifying transcripts at steady state with standard mRNA-seq. It is possible that we observed a repressive effect of CHG hypermethylation on a subset of genes because we profiled a developmental stage in which hundreds of genes become transcriptionally activated at the onset of embryo maturation. Nevertheless, we cannot completely exclude that the repression of hypermethylated genes undergoing transcriptional activation is due to secondary effects of other genes influenced by CMT3-induced hypermethylation. Importantly, CMT3-induced CHG hypermethylation due at least partially to loss of miR823 repression in embryos is largely maintained for weeks after detectable miRNA activity (Figure 3). Therefore, epigenetic collateral damage occurring in embryos may also negatively impact gene expression later in life. However, additional experiments are required to directly test the relationship between CMT3-induced hypermethylation and gene expression.

Transcriptional (Ning et al., 2020), post-transcriptional (Figure 3), post-translational (Deng et al., 2016), post-hoc (Saze et al., 2008) and perhaps substrate-related (Figure 4) mechanisms fine-tune CMT3 activities to levels required to specifically silence mutagenic TEs but not genes. However, errors in restricting CMT3 to heterochromatin are inevitable on an evolutionary timescale (Zhang et al., 2020) and recent studies indicate that CMT3-induced methylation of genes precedes gbM (Wendte et al., 2019). Because gbM can be stably maintained over many generations by MET1 and its functional significance is debatable (Bewick et al., 2016; Bewick et al., 2019; Choi et al., 2020; Coleman-Derr and Zilberman, 2012; Le et al., 2020; Picard and Gehring, 2017; Shahzad et al., 2021; Takuno and Gaut, 2013; Wendte et al., 2019; Williams et al., 2021; Zilberman, 2017), it cannot be excluded that gbM is merely an evolutionary record of epigenetic collateral damage events that occurred in the past (Bewick and Schmitz, 2017; Bewick et al., 2017). Our results suggest that derepressed CMT3 and MET1 both prefer genic regions characterized by increased nucleosome stability (Figure 4). Accordingly, CMT3-induced CHG hypermethylation tends to occur away from transcription start and end sites of genes in a nearly identical pattern as observed for gbM (Figure 4). We propose that CHG methylation is more tolerated in central/3' biased regions because they are relatively inaccessible to trans-acting factors that regulate transcription. Moreover, our results tentatively suggest that CMT3-induced hypermethylation can repress genes that are transcriptionally activated (Figure 5). Perhaps genes that are consistently expressed can accumulate CHG methylation without having a large effect on steady state transcript levels and resulting fitness penalties, and thus be more likely to accumulate gbM over evolutionary time. In other words, miR823-mediated repression is one of several ways to prevent CMT3 from ectopically methylating protein-coding genes. However, CMT3 off-targeting on genes may still occur despite these complex regulatory mechanisms and the resulting epigenetic collateral damage can be recorded as heritable gbM. The characteristic features of gbM may not pertain to its current

functions, but rather the consequences of transient CHG methylation that occurred in the past and were selected on during evolution.

Materials and methods

Key resources table

Reagent type (species) or resource	Designation	Source or reference	Identifiers	Additional information
Gene (Arabidopsis thaliana)	CHROMOMETHYLASE 3 (CMT3)	TAIR	AT1G69770	
Gene (Arabidopsis thaliana)	MICRORNA 823A (MIR823A)	TAIR	AT3G13724	
Genetic reagent (Arabidopsis thaliana)	miR823-cleavable CMT3 (cCMT3)	this paper		pAlligatorR43/promoterCMT3::genomicCMT3
Genetic reagent (Arabidopsis thaliana)	miR823-resistant CMT3 (rCMT3)	this paper		pAlligatorR43/promoterCMT3::resistantCMT3 (generated from cCMT3 with site-directed mutagenesis)
Recombinant DNA reagent	pAlligatorR43 (plasmid)	DOI:10.7554/eLife.04501		mCherry selection marker
Recombinant DNA reagent	pHSE401 (plasmid)	Addgene	#62201	CRISPR/Cas9 plasmid
Recombinant DNA reagent	pCBCD-T1T2 (plasmid)	Addgene	#50590	CRISPR/Cas9 plasmid
Strain (Arabidopsis thaliana)	cmt3-11T	NASC	SALK_148381	T-DNA insertion mutant of CMT3
Strain (Arabidopsis thaliana)	mir823-1	this paper		miR823 knockout mutant
Strain (Arabidopsis thaliana)	mir823-2	this paper		miR823 knockout mutant
Commercial kit	Q5 Site-Directed Mutagenesis Kit	New England Biolabs	#E0554S	
Commercial kit	Fast SYBR Green Master Mix	Roche	#06402712001	
Commercial kit	SuperScript III Reverse Transcriptase	Thermo Fisher Scientific	#18080093	
Commercial kit	TRIzol	Invitrogen	#15596026	
Software	Lightcycler 96	Roche Diagnostics	Version 1.1.0.1320	

Plant material and growth conditions

Arabidopsis thaliana accession Columbia-0 (Col-0) were grown in controlled growth chambers at 20–22°C under a 16 hr light/8 hr dark cycle with incandescent lights (130 to 150 μmol/m²/s).

Generation of transgenic lines

The control genomic CMT3 construct (miR823-cleavable; cCMT3) was generated by PCR amplification of the CMT3 locus including 1408 bp upstream and 730 bp downstream of the TAIR10-annotated transcription start and end sites, respectively. PCR primers included overhangs for subsequent Gibson assembly into MultiSite-Gateway destination vector pAlligatorR43 (Kawashima et al., 2013). The miR823-resistant CMT3 construct (rCMT3) was generated by PCR site-directed mutagenesis (Q5 Site-Directed Mutagenesis Kit, New England Biolabs) using the cCMT3 construct as a template to introduce six silent mutations as shown in Figure 3—figure supplement 1C. Both cCMT3 and rCMT3 construct sequences were analyzed for mutations using Sanger sequencing. All primers used are listed in the Supplementary file 5. The constructs were transformed into cmt3-11T (SALK_148381) using the Agrobacterium floral dip method (Clough and Bent, 1998), and

transformants were selected based on seed-coat RFP signal under fluorescent light (Zeiss SteREO DiscoveryV.8). Multiple independent first-generation transgenic (T1) lines were identified for cCMT3 and rCMT3, and three and four were characterized in bent cotyledon embryos for each, respectively.

Generation of CRISPR/Cas9 knockout mutants for *MIR823*

CRISPR/Cas9 knockout mutants in *MIR823* were created by using a modified pHSE401 binary vector (Addgene #62201) according to the protocol detailed by Xing *et al.*, 2014. Primers containing the sequences for the two guide RNAs targeting the *MIR823* locus flanking the miR823 sequence (Figure 3—figure supplement 1A and Supplementary file 5) were amplified together with the pCBCD-T1T2 plasmid (Addgene #50590), and the resulting PCR product was subsequently assembled into the pHSE401 binary vector using GoldenGate cloning method (Xing *et al.*, 2014). Plants were transformed with the floral dip method as described above; and Cas9-positive seeds were selected based on the presence of seed coat RFP signal. Deletion lines were identified with PCR using primers flanking gRNA-targeted sites (Figure 3—figure supplement 1A and Supplementary file 5). Deletion mutants were confirmed and mapped by Sanger sequencing.

qRT-PCR analysis

Leaves (two-week old rosettes), floral clusters (five weeks) and bent cotyledon embryos (eight DAP) were homogenized in 500 µl TRIzol reagent (Invitrogen) and total RNA was isolated and purified according to manufacturer's recommendations. For mRNA, 200 ng of total RNA was used for cDNA synthesis with SuperScript III Reverse Transcriptase (Thermo Fisher Scientific). The cDNA was diluted two-fold for embryos or ten-fold for leaves and floral buds with nuclease-free water. Two µL of diluted cDNA was used as a template for the qRT-PCR with Fast SYBR Green Master Mix (Roche) on a LightCycler 96 instrument (Roche) with two technical replicates for each biorep. For miRNA823 quantification, corresponding stem-loop primers were added to the RT reaction (adapted from Yang *et al.*, 2014) and miR823 levels were measured using Fast SYBR Green Master Mix (Roche) with miRNA823-specific forward primer and a stem-loop specific universal reverse primer. U6 snRNA was used as the reference RNA (adapted from Shen *et al.*, 2010). Primers used for qRT-PCR are listed in Supplementary file 5.

Sample size estimation, embryo isolation, and nucleic acid extraction

Sample sizes were determined based on a combination of the required statistical power, ability to acquire samples and cost of the experiments. Bent cotyledon embryos were dissected from seeds 8 days after pollination and also selected based on morphology to ensure accurate staging. Embryos were serially washed 4× with nuclease-free water under an inverted microscope. Approximately 50 embryos per replicate were isolated and stored at −80°C until further use. RNA was isolated as previously described (Lutzmayr *et al.*, 2017; Plotnikova *et al.*, 2019). Genomic DNA was extracted from embryos and 3-week old plants using Quick-DNA Micro prep Kit (Zymo D3020) according to the recommendations of the manufacturer.

DNA methylation profiling and analysis

MethylC-Seq libraries were generated as described previously (Papareddy *et al.*, 2020) and sequenced in single-read mode on an Illumina HiSeq 2500 or Nextseq 550 instrument. Adapters and the first six bases corresponding to random hexamers used during the pre-amplification step were trimmed from MethylC-seq reads using Trim Galore. Bisulfite-converted reads were aligned against the TAIR10 genome (Lamesch *et al.*, 2012) in non-directional mode using Bismark (bismark -non-directional -q -score-min L,0,-0.4) (Krueger and Andrews, 2011). Methylpy software was used to extract weighted methylation rates for each available cytosine from BAM files containing only deduplicated and uniquely mapped reads (Schultz *et al.*, 2015). Reads mapping to the unmetabolized chloroplast genome were used to calculate bisulfite conversion rates. FASTQ files obtained from publicly available methylomes generated from sperm (Ibarra *et al.*, 2012), early torpedo (Pignatta *et al.*, 2015), mid-torpedo to early maturation (Hsieh *et al.*, 2009), mature green embryos (Bouyer *et al.*, 2017) and DNA methylation mutant leaves (Stroud *et al.*, 2013) were also processed in a similar manner except that alignments were performed in directional mode and only 5' end

nucleotides of the reads with m-bias were removed. Differentially methylated regions (DMRs) were identified using *Methylpy* (Schultz et al., 2015). Briefly, two biological replicates were pooled and differentially methylated cytosines (DMCs) were identified by root mean squared tests with false discovery rates ≤ 0.01 . DMRs were defined by collapsing DMCs with ≥ 4 reads within 500 bps to single units requiring ≥ 8 and ≥ 4 DMCs for CG and CHN sites, respectively ($N = A, T, C, G$; $H \neq G$). Using these parameters, DMRs were identified across floral bud, early heart, early torpedo, bent cotyledon, mature green and leaf samples, and merged into a single bedFile using the BEDtools *merge* function (Quinlan and Hall, 2010). Resulting DMRs were then used to calculate the methylation rate on all analyzed tissues and genotypes. We assigned that a gene and a DMR are associated if the DMR is overlapping within 1.5 kb upstream or downstream of TAIR10 annotated gene bodies using BEDtools *closest* function. For down-regulated genes overlapping with DMRs with above criteria, significance was tested using BEDtools *fisher* function with nuclear genome as a background control.

mRNA profiling and analysis

Smart-seq2 mRNA libraries were generated from 1 μ l of the 7 μ l bent cotyledon embryo total RNA as previously described (Hofmann et al., 2019; Picelli et al., 2014). Both amplified cDNA and final libraries were inspected using Agilent HS NGS Fragment Kit (DNF-474) to control for library quality and proper length distributions. Libraries were sequenced in single-read mode on an Illumina HiSeq 2500 or NextSeq 550 machine. Raw FASTQ files from technical replicates were merged, quality filtered and trimmed for adapter sequences with *Trim Galore* using default parameters. Trimmed reads were aligned using STAR (Dobin et al., 2013) against a genome index generated using the TAIR10 genome fasta file and all transcripts in the GTF of Ensembl build TAIR10 annotation set (release version 44). Aligned transcriptome bam files were used to quantify read counts per gene and transcript abundance using RSEM (Li and Dewey, 2011). Along with the transcriptomes generated in this study, publicly available embryonic transcriptomes Hofmann et al., 2019 used for PCA were analyzed in the same fashion as described above (Supplementary file 6). Prior to PCA (Figure 5A), read counts derived from nuclear protein-coding genes were subjected to variance stabilizing transformations using DESeq2 (Love et al., 2014). Differential gene expression analysis was performed using DESeq2 for genes with at least five aligned reads. Genes with ≥ 2 -fold differences and adjusted p-value ≤ 0.01 were classified as differentially expressed genes (DEGs). Nearest-neighbor genes in Figure 1A,B were classified based on Euclidean distance. First, the centroid expression of MET1 and VIM1/2/3 was calculated for all tissue types represented in the developmental time series. This centroid value was then used to calculate Euclidean distance of all TAIR10-annotated protein-coding genes and sorted based on their distances.

ChIP-seq analysis

ChIP-seq data for H2A variants and H3K9me2 were downloaded from GSE50942 (Yelagandula et al., 2014) and GSE51304 (Stroud et al., 2014), respectively. H3K9 acetylation marks were from GSE98214 (Wang et al., 2019). H3K4me3 marks were obtained from GSE152243 (Maher, 2020). All FASTQ files were trimmed and quality filtered using *Trim Galore* default parameters. Trimmed reads were aligned against the TAIR10 genome using BWA-MEM (Li and Durbin, 2009). Multi-mapping reads and clonal duplicates were removed using *MarkDuplicates* from the Picard Tools suite (Toolkit, 2019). The resulting BAM files containing alignments were sorted, indexed and used as input for the *bamCoverage* function of deepTools (Ramírez et al., 2014) to obtain genome normalized coverage with parameters *-normalizeUsing 'RPGC'*. Processed bigwig files for H1 Chromatin Affinity purification followed by sequencing (ChAP) and DNase-seq datasets were obtained from GSE122394 (Choi et al., 2020). MNase-Seq data was obtained from GSE113556 (Rutowicz et al., 2019). ATAC-seq processed bigwig files for WT and *ddm1* mutants were from GSE155503 (Zhong et al., 2021).

Metaplots

ChIP, ATAC, MNase, DNase, and MethylC-seq metaplots were plotted using the R library *Seqplots* (Stempor and Ahringer, 2016). Body, upstream, and downstream regions of TEs or genes were split into equal-sized bins, and the average levels for each bin was calculated and plotted.

CMT3 transgene copy number estimation

CMT3 transgene copy number was estimated using two methods: qPCR and coverage calculation. For the qPCR method, genomic DNA was extracted from leaves of three-week old plants using the CTAB DNA isolation method (Aboul-Maaty and Oraby, 2019). Relative transgene copy number was determined by using the qPCR-based method as described (Shepherd et al., 2009). *ACTIN2* was used as a control gene while transgene copy number was calculated based on CMT3 levels. For the coverage method, Bismark-aligned and deduplicated BAM files from WT, cCMT3 and rCMT3 lines were processed with DeepTools to obtain normalized genome coverage as bins per million mapped reads (BPM) units with the `bamCoverage` function and following parameters: `–binsize 50 –skipNAs –normalizeUsing ‘BPM’ –ignoreForNormalization mitochondria chloroplast`. The resulting bigwig files were used to calculate genome-wide coverage fold-changes relative to WT using the deepTools function `bigwigCompare –skipNAs –operation ‘ratio’`. CMT3 locus was displayed with the Integrative Genomics Viewer (IGV).

Availability of data and material

All sequencing data generated in this study are available at the National Center for Biotechnology Information Gene Expression Omnibus (NCBI GEO, <https://www.ncbi.nlm.nih.gov/geo/>) under accession number GSE171198. ChIP-Seq and mRNA-seq bioinformatic analysis pipelines were based on Nextflow (Di Tommaso et al., 2017) and the nf-core framework (Ewels et al., 2020) are available at <https://github.com/Gregor-Mendel-Institute/RKP2021-CMT3>, [copy archived at [swh:1:rev:89d7e2ea78af1969bb161640baed09296ed2485f](https://www.swh.io/rev/89d7e2ea78af1969bb161640baed09296ed2485f) (Papareddy, 2021)].

Acknowledgements

We thank the Vienna Biocenter Core Facilities GmbH (VBCF) Next Generation Sequencing and Plant Sciences Facilities for next-generation sequencing and plant growth chamber access, respectively, and the Institute of Molecular Pathology-Institute of Molecular Biology-Gregor Mendel Institute Molecular Biology Services for instrument access and support. RKP personally thanks Pierre Bourguet and Michael Borg for invaluable discussions. We also thank Zdravko Lorkovic, Bhagyshree Jamge, Robin Burns, Eriko Sasaki, Magnus Nordborg and Frédéric Berger for sharing thoughts and reagents; and members of the Nodine lab for valuable input. This work was supported by the European Research Council under the European Union’s Horizon 2020 Research and Innovation Program grant 637888 to MDN.

Additional information

Funding

Funder	Grant reference number	Author
H2020 European Research Council	637888	Michael D Nodine

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

Ranjith K Papareddy, Conceptualization, Software, Formal analysis, Investigation, Methodology, Writing - original draft, Writing - review and editing; Katalin Páldi, Investigation, Methodology; Anna D Smolka, Investigation; Patrick Hüther, Software; Claude Becker, Supervision; Michael D Nodine, Conceptualization, Supervision, Funding acquisition, Writing - original draft, Writing - review and editing

Author ORCIDs

Claude Becker  <http://orcid.org/0000-0003-3406-4670>

Michael D Nodine  <https://orcid.org/0000-0002-6204-8857>

Decision letter and Author response

Decision letter <https://doi.org/10.7554/eLife.69396.sa1>

Author response <https://doi.org/10.7554/eLife.69396.sa2>

Additional files

Supplementary files

- Supplementary file 1. CG differentially methylated regions during development.
- Supplementary file 2. CHG differentially methylated regions during development.
- Supplementary file 3. Transcript levels in rCMT3 compared to wild-type embryos.
- Supplementary file 4. CHG differentially methylated regions in rCMT3 compared to wild type.
- Supplementary file 5. Oligonucleotides used in this study.
- Supplementary file 6. MethylC-seq and mRNA-seq mapping statistics.
- Transparent reporting form

Data availability

All sequencing data generated in this study are publicly available at the National Center for Biotechnology Information Gene Expression Omnibus (NCBI GEO, <https://www.ncbi.nlm.nih.gov/geo/>) under accession number GSE171198. ChIP-Seq and mRNA-seq bioinformatic analysis pipelines were based on Nextflow and the nf-core framework are available at <https://github.com/Gregor-Mendel-Institute/RKP2021-CMT3> (copy archived at <https://archive.softwareheritage.org/wh:1:rev:89d7e2ea78af1969bb161640baed09296ed2485f>).

The following dataset was generated:

Author(s)	Year	Dataset title	Dataset URL	Database and Identifier
Nodine M, Papareddy R	2021	Repression of CHROMOMETHYLASE 3 Prevents Epigenetic Collateral Damage in Arabidopsis	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE171198	NCBI Gene Expression Omnibus, GSE171198

The following previously published datasets were used:

Author(s)	Year	Dataset title	Dataset URL	Database and Identifier
Nodine M, Hofmann F, Schon M	2018	The embryonic transcriptome of Arabidopsis thaliana	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE121236	NCBI Gene Expression Omnibus, GSE121236
Nodine M, Papareddy R	2020	Chromatin regulates expression of small RNAs to help maintain transposon methylome homeostasis in Arabidopsis	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE152971	NCBI Gene Expression Omnibus, GSE152971
Nishimura T, Zilberman D, Nishimura T, Zilberman D	2012	Active DNA demethylation in plant companion cells reinforces transposon methylation in gametes	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE38935	NCBI Gene Expression Omnibus, GSE38935
Gehring	2014	Natural epigenetic polymorphisms lead to intraspecific variation in Arabidopsis gene imprinting	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE52814	NCBI Gene Expression Omnibus, GSE52814
Hsieh T, Ibarra CA, Silva P, Zemach A, Williams LE, Fisher RL, Zilberman D	2009	Genome-wide demethylation of Arabidopsis endosperm	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE15922	NCBI Gene Expression Omnibus, GSE15922
Bouyer D, Roudier F, Colot V, Bouyer D, Roudier F, Colot V	2017	DNA methylation dynamics during early plant life	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE85975	NCBI Gene Expression Omnibus, GSE85975
Stroud H	2013	Comprehensive analysis of	https://www.ncbi.nlm.nih.gov/geo/	NCBI Gene

		silencing mutants reveals complex regulation of the <i>Arabidopsis</i> methylome	nih.gov/geo/query/acc.cgi?acc=GSE39901	Expression Omnibus, GSE39901
Yelagandula R, Stroud H	2014	The Histone Variant H2A.W Defines Heterochromatin and Promotes Chromatin Condensation in <i>Arabidopsis</i>	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE50942	NCBI Gene Expression Omnibus, GSE50942
Stroud H	2013	Non-CG methylation patterns shape the epigenetic landscape in <i>Arabidopsis</i>	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE51304	NCBI Gene Expression Omnibus, GSE51304
Wang L, Wang C, Liu X, Cheng J, Li S, Zhu J, Gong Z	2018	Peroxisomal β -oxidation regulates histone acetylation and DNA methylation in <i>Arabidopsis</i>	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE98214	NCBI Gene Expression Omnibus, GSE98214
Mahe KA, Wang D, Deal RB, Mahe KA, Wang D, Deal RB	2020	Differences in directionality of RNA polymerase initiation underlie epigenome differences between plants and animals	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE152243	NCBI Gene Expression Omnibus, GSE152243
Choi J, Lyons DB, Kim MY, Choi J, Lyons DB, Kim MY	2019	DNA methylation and histone H1 jointly repress transposable elements and aberrant intragenic transcripts	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE122394	NCBI Gene Expression Omnibus, GSE122394
Lirski M, Kroten MA, Lirski M, Kroten MA	2019	MNase analysis of linker histone H1 mutant	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE113556	NCBI Gene Expression Omnibus, GSE113556
Zhenhui Z	2021	DNA methylation-linked chromatin accessibility affects genomic architecture in <i>Arabidopsis</i>	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE155503	NCBI Gene Expression Omnibus, GSE155503

References

- Aboul-Maaty NA-F, Oraby HA-S. 2019. Extraction of high-quality genomic DNA from different plant orders applying a modified CTAB-based method. *Bulletin of the National Research Centre* **43**:25. DOI: <https://doi.org/10.1186/s42269-019-0066-1>
- Antunez-Sanchez J, Naish M, Ramirez-Prado JS, Ohno S, Huang Y, Dawson A, Opassathian K, Manza-Mianza D, Ariel F, Raynaud C, Wibowo A, Daron J, Ueda M, Latrasse D, Slotkin RK, Weigel D, Benhamed M, Gutierrez-Marcos J. 2020. A new role for histone demethylases in the maintenance of plant genome integrity. *eLife* **9**: e58533. DOI: <https://doi.org/10.7554/eLife.58533>, PMID: 33107825
- Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**:796–815. DOI: <https://doi.org/10.1038/35048692>, PMID: 11130711
- Bewick AJ, Ji L, Niederhuth CE, Willing EM, Hofmeister BT, Shi X, Wang L, Lu Z, Rohr NA, Hartwig B, Kiefer C, Deal RB, Schmutz J, Grimwood J, Stroud H, Jacobsen SE, Schneeberger K, Zhang X, Schmitz RJ. 2016. On the origin and evolutionary consequences of gene body DNA methylation. *PNAS* **113**:9111–9116. DOI: <https://doi.org/10.1073/pnas.1604666113>, PMID: 27457936
- Bewick AJ, Niederhuth CE, Ji L, Rohr NA, Griffin PT, Leebens-Mack J, Schmitz RJ. 2017. The evolution of CHROMOMETHYLASES and gene body DNA methylation in plants. *Genome Biology* **18**:65. DOI: <https://doi.org/10.1186/s13059-017-1195-1>, PMID: 28457232
- Bewick AJ, Zhang Y, Wendte JM, Zhang X, Schmitz RJ. 2019. Evolutionary and experimental loss of gene body methylation and its consequence to gene expression. *G3: Genes, Genomes, Genetics* **9**:2441–2445. DOI: <https://doi.org/10.1534/g3.119.400365>, PMID: 31147389
- Bewick AJ, Schmitz RJ. 2017. Gene body DNA methylation in plants. *Current Opinion in Plant Biology* **36**:103–110. DOI: <https://doi.org/10.1016/j.pbi.2016.12.007>, PMID: 28258985
- Blevins T, Podicheti R, Mishra V, Marasco M, Wang J, Rusch D, Tang H, Pikaard CS. 2015. Identification of pol IV and RDR2-dependent precursors of 24 nt siRNAs guiding de novo DNA methylation in *Arabidopsis*. *eLife* **4**: e09591. DOI: <https://doi.org/10.7554/eLife.09591>, PMID: 26430765
- Borges F, Donoghue MTA, LeBlanc C, Wear EE, Tanurdžić M, Berube B, Brooks A, Thompson WF, Hanley-Bowdoin L, Martienssen RA. 2021. Loss of Small-RNA-Directed DNA methylation in the plant cell cycle promotes germline reprogramming and somaclonal variation. *Current Biology* **31**:591–600. DOI: <https://doi.org/10.1016/j.cub.2020.10.098>, PMID: 33275892
- Bourguet P, Picard CL, Yelagandula R, Pélissier T, Lorković ZJ, Feng S, Pouch-Pélissier M-N, Schmücker A, Jacobsen SE, Berger F. 2021. The histone variant H2A.W and Linker histone H1 co-regulate heterochromatin accessibility and DNA methylation. *Nat Commun* **11**:2683. DOI: <https://doi.org/10.1038/s41467-021-22993>
- Bouyer D, Kramdi A, Kassam M, Heese M, Schnittger A, Roudier F, Colot V. 2017. DNA methylation dynamics during early plant life. *Genome Biology* **18**:179. DOI: <https://doi.org/10.1186/s13059-017-1313-0>, PMID: 28942733

- Calarco JP, Borges F, Donoghue MT, Van Ex F, Jullien PE, Lopes T, Gardner R, Berger F, Feijó JA, Becker JD, Martienssen RA. 2012. Reprogramming of DNA methylation in pollen guides epigenetic inheritance via small RNA. *Cell* **151**:194–205. DOI: <https://doi.org/10.1016/j.cell.2012.09.001>, PMID: 23000270
- Cao X, Jacobsen SE. 2002. Locus-specific control of asymmetric and CpNpG methylation by the DRM and CMT3 methyltransferase genes. *PNAS* **99**:16491–16498. DOI: <https://doi.org/10.1073/pnas.162371599>, PMID: 12151602
- Chodavarapu RK, Feng S, Bernatavichute YV, Chen PY, Stroud H, Yu Y, Hetzel JA, Kuo F, Kim J, Cokus SJ, Casero D, Bernal M, Huijser P, Clark AT, Krämer U, Merchant SS, Zhang X, Jacobsen SE, Pellegrini M. 2010. Relationship between nucleosome positioning and DNA methylation. *Nature* **466**:388–392. DOI: <https://doi.org/10.1038/nature09147>, PMID: 20512117
- Choi J, Lyons DB, Kim MY, Moore JD, Zilberman D. 2020. DNA methylation and histone H1 jointly repress transposable elements and aberrant intragenic transcripts. *Molecular Cell* **77**:310–323. DOI: <https://doi.org/10.1016/j.molcel.2019.10.011>, PMID: 31732458
- Clough SJ, Bent AF. 1998. Floral dip: a simplified method for *Agrobacterium*-mediated transformation of *Arabidopsis thaliana*: floral dip transformation of *Arabidopsis*. *The Plant Journal : For Cell and Molecular Biology* **16**:735–743. DOI: <https://doi.org/10.1046/j.1365-313x.1998.00343>
- Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, Pradhan S, Nelson SF, Pellegrini M, Jacobsen SE. 2008. Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. *Nature* **452**:215–219. DOI: <https://doi.org/10.1038/nature06745>, PMID: 18278030
- Coleman-Derr D, Zilberman D. 2012. Deposition of histone variant H2A.Z within gene bodies regulates responsive genes. *PLOS Genetics* **8**:e1002988. DOI: <https://doi.org/10.1371/journal.pgen.1002988>, PMID: 23071449
- Deng S, Jang I-C, Su L, Xu J, Chua N-H. 2016. JM24 targets CHROMOMETHYLASE3 for proteasomal degradation in *Arabidopsis*. *Genes & Development* **30**:251–256. DOI: <https://doi.org/10.1101/gad.274647.115>
- Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. 2017. Nextflow enables reproducible computational workflows. *Nature Biotechnology* **35**:316–319. DOI: <https://doi.org/10.1038/nbt.3820>, PMID: 28398311
- Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**:15–21. DOI: <https://doi.org/10.1093/bioinformatics/bts635>
- Du J, Zhong X, Bernatavichute YV, Stroud H, Feng S, Caro E, Vashisht AA, Terragni J, Chin HG, Tu A, Hetzel J, Wohlschlegel JA, Pradhan S, Patel DJ, Jacobsen SE. 2012. Dual binding of chromomethylase domains to H3K9me2-containing nucleosomes directs DNA methylation in plants. *Cell* **151**:167–180. DOI: <https://doi.org/10.1016/j.cell.2012.07.034>
- Du J, Johnson LM, Groth M, Feng S, Hale CJ, Li S, Vashisht AA, Wohlschlegel JA, Patel DJ, Jacobsen SE. 2014. Mechanism of DNA methylation-directed histone methylation by KRYPTONITE. *Molecular Cell* **55**:495–504. DOI: <https://doi.org/10.1016/j.molcel.2014.06.009>, PMID: 25018018
- Du J, Johnson LM, Jacobsen SE, Patel DJ. 2015. DNA methylation pathways and their crosstalk with histone methylation. *Nature Reviews Molecular Cell Biology* **16**:519–532. DOI: <https://doi.org/10.1038/nrm4043>, PMID: 26296162
- Ewels PA, Peltzer A, Fillinger S, Patel H, Alneberg J, Wilm A, Garcia MU, Di Tommaso P, Nahnsen S. 2020. The nf-core framework for community-curated bioinformatics pipelines. *Nature Biotechnology* **38**:276–278. DOI: <https://doi.org/10.1038/s41587-020-0439-x>, PMID: 32055031
- Fan Y, Nikitina T, Morin-Kensicki EM, Zhao J, Magnuson TR, Woodcock CL, Skoultschi AI. 2003. H1 linker histones are essential for mouse development and affect nucleosome spacing in vivo. *Molecular and Cellular Biology* **23**:4559–4572. DOI: <https://doi.org/10.1128/MCB.23.13.4559-4572.2003>
- Feng S, Cokus SJ, Zhang X, Chen PY, Bostick M, Goll MG, Hetzel J, Jain J, Strauss SH, Halpern ME, Ukomadu C, Sadler KC, Pradhan S, Pellegrini M, Jacobsen SE. 2010. Conservation and divergence of methylation patterning in plants and animals. *PNAS* **107**:8689–8694. DOI: <https://doi.org/10.1073/pnas.1002720107>, PMID: 20395551
- Finnegan EJ, Dennis ES. 1993. Isolation and identification by sequence homology of a putative cytosine methyltransferase from *Arabidopsis thaliana*. *Nucleic Acids Research* **21**:2383–2388. DOI: <https://doi.org/10.1093/nar/21.10.2383>, PMID: 8389441
- Gouil Q, Baulcombe DC. 2016. DNA methylation signatures of the plant chromomethyltransferases. *PLOS Genetics* **12**:e1006526. DOI: <https://doi.org/10.1371/journal.pgen.1006526>, PMID: 27997534
- Gutzat R, Rembart K, Nussbaumer T, Hofmann F, Pisupati R, Bradamante G, Daubel N, Gaidora A, Lettner N, Donà M, Nordborg M, Nodine M, Mittelsten Scheid O. 2020. *Arabidopsis* shoot stem cells display dynamic transcription and DNA methylation patterns. *The EMBO Journal* **39**:e103667. DOI: <https://doi.org/10.15252/emboj.2019103667>, PMID: 32815560
- Henderson IR, Jacobsen SE. 2007. Epigenetic inheritance in plants. *Nature* **447**:418–424. DOI: <https://doi.org/10.1038/nature05917>, PMID: 17522675
- Henderson IR, Jacobsen SE. 2008. Tandem repeats upstream of the *Arabidopsis* endogene SDC recruit non-CG DNA methylation and initiate siRNA spreading. *Genes & Development* **22**:1597–1606. DOI: <https://doi.org/10.1101/gad.1667808>, PMID: 18559476
- Herr AJ, Jensen MB, Dalmay T, Baulcombe DC. 2005. RNA polymerase IV directs silencing of endogenous DNA. *Science* **308**:118–120. DOI: <https://doi.org/10.1126/science.1106910>, PMID: 15692015
- Hofmann F, Schon MA, Nodine MD. 2019. The embryonic transcriptome of *Arabidopsis thaliana*. *Plant Reproduction* **32**:77–91. DOI: <https://doi.org/10.1007/s00497-018-00357-2>, PMID: 30610360

- Hsieh TF, Ibarra CA, Silva P, Zemach A, Eshed-Williams L, Fischer RL, Zilberman D. 2009. Genome-wide demethylation of Arabidopsis endosperm. *Science* **324**:1451–1454. DOI: <https://doi.org/10.1126/science.1172417>, PMID: 19520962
- Ibarra CA, Feng X, Schoft VK, Hsieh TF, Uzawa R, Rodrigues JA, Zemach A, Chumak N, Machlicova A, Nishimura T, Rojas D, Fischer RL, Tamaru H, Zilberman D. 2012. Active DNA demethylation in plant companion cells reinforces transposon methylation in gametes. *Science* **337**:1360–1364. DOI: <https://doi.org/10.1126/science.1224839>, PMID: 22984074
- Inagaki S, Miura-Kamio A, Nakamura Y, Lu F, Cui X, Cao X, Kimura H, Saze H, Kakutani T. 2010. Autocatalytic differentiation of epigenetic modifications within the Arabidopsis genome. *The EMBO Journal* **29**:3496–3506. DOI: <https://doi.org/10.1038/emboj.2010.227>, PMID: 20834229
- Inagaki S, Takahashi M, Hosaka A, Ito T, Toyoda A, Fujiyama A, Tarutani Y, Kakutani T. 2017. Gene-body chromatin modification dynamics mediate epigenome differentiation in Arabidopsis. *The EMBO Journal* **36**:970–980. DOI: <https://doi.org/10.15252/embj.201694983>, PMID: 28100676
- Ito T, Tarutani Y, To TK, Kassam M, Duvernois-Berthet E, Cortijo S, Takashima K, Saze H, Toyoda A, Fujiyama A, Colot V, Kakutani T. 2015. Genome-wide negative feedback drives transgenerational DNA methylation dynamics in Arabidopsis. *PLOS Genetics* **11**:e1005154. DOI: <https://doi.org/10.1371/journal.pgen.1005154>, PMID: 25902052
- Jackson JP, Lindroth AM, Cao X, Jacobsen SE. 2002. Control of CpNpG DNA methylation by the KRYPTONITE histone H3 methyltransferase. *Nature* **416**:556–560. DOI: <https://doi.org/10.1038/nature731>, PMID: 11898023
- Jacobsen SE, Meyerowitz EM. 1997. Hypermethylated SUPERMAN epigenetic alleles in Arabidopsis. *Science* **277**:1100–1103. DOI: <https://doi.org/10.1126/science.277.5329.1100>, PMID: 9262479
- Ji L, Mathioni SM, Johnson S, Tucker D, Bewick AJ, Do Kim K, Daron J, Slotkin RK, Jackson SA, Parrott WA, Meyers BC, Schmitz RJ. 2019. Genome-Wide reinforcement of DNA methylation occurs during somatic embryogenesis in soybean. *The Plant Cell* **31**:2315–2331. DOI: <https://doi.org/10.1105/tpc.19.00255>, PMID: 31439802
- Kato M, Miura A, Bender J, Jacobsen SE, Kakutani T. 2003. Role of CG and non-CG methylation in immobilization of transposons in Arabidopsis. *Current Biology* **13**:421–426. DOI: [https://doi.org/10.1016/S0960-9822\(03\)00106-4](https://doi.org/10.1016/S0960-9822(03)00106-4), PMID: 12620192
- Kawakatsu T, Stuart T, Valdes M, Breakfield N, Schmitz RJ, Nery JR, Ulrich MA, Han X, Lister R, Benfey PN, Ecker JR. 2016. Unique cell-type-specific patterns of DNA methylation in the root meristem. *Nature Plants* **2**:16058. DOI: <https://doi.org/10.1038/nplants.2016.58>, PMID: 27243651
- Kawakatsu T, Nery JR, Castanon R, Ecker JR. 2017. Dynamic DNA methylation reconfiguration during seed development and germination. *Genome Biology* **18**:171. DOI: <https://doi.org/10.1186/s13059-017-1251-x>
- Kawashima T, Kitamura K, Suzuki K, Nonaka M, Kamijo S, Takemoto-Kimura S, Kano M, Okuno H, Ohki K, Bito H. 2013. Functional labeling of neurons and their projections using the synthetic activity-dependent promoter E-SARE. *Nature Methods* **10**:889–895. DOI: <https://doi.org/10.1038/nmeth.2559>, PMID: 23852453
- Krueger F, Andrews SR. 2011. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**:1571–1572. DOI: <https://doi.org/10.1093/bioinformatics/btr167>, PMID: 21493656
- Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M, Karthikeyan AS, Lee CH, Nelson WD, Ploetz L, Singh S, Wensel A, Huala E. 2012. The Arabidopsis information resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Research* **40**:D1202–D1210. DOI: <https://doi.org/10.1093/nar/gkr1090>, PMID: 22140109
- Law JA, Jacobsen SE. 2010. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nature Reviews Genetics* **11**:204–220. DOI: <https://doi.org/10.1038/nrg2719>, PMID: 20142834
- Le NT, Harukawa Y, Miura S, Boer D, Kawabe A, Saze H. 2020. Epigenetic regulation of spurious transcription initiation in Arabidopsis. *Nature Communications* **11**:3224. DOI: <https://doi.org/10.1038/s41467-020-16951-w>, PMID: 32591528
- Lee YS, Maple R, Dürr J, Dawson A, Tamim S, Del Genio C, Papareddy R, Luo A, Lamb JC, Amantia S, Sylvester AW, Birchler JA, Meyers BC, Nodine MD, Rouster J, Gutierrez-Marcos J. 2021. A transposon surveillance mechanism that safeguards plant male fertility during stress. *Nature Plants* **7**:34–41. DOI: <https://doi.org/10.1038/s41477-020-00818-5>, PMID: 33398155
- Li X, Harris CJ, Zhong Z, Chen W, Liu R, Jia B, Wang Z, Li S, Jacobsen SE, Du J. 2018. Mechanistic insights into plant SUVH family H3K9 methyltransferases and their binding to context-biased non-CG DNA methylation. *PNAS* **115**:E8793–E8802. DOI: <https://doi.org/10.1073/pnas.1809841115>, PMID: 30150382
- Li B, Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**:323. DOI: <https://doi.org/10.1186/1471-2105-12-323>, PMID: 21816040
- Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**:1754–1760. DOI: <https://doi.org/10.1093/bioinformatics/btp324>, PMID: 19451168
- Lin JY, Le BH, Chen M, Henry KF, Hur J, Hsieh TF, Chen PY, Pelletier JM, Pellegrini M, Fischer RL, Harada JJ, Goldberg RB. 2017. Similarity between soybean and Arabidopsis seed methylomes and loss of non-CG methylation does not affect seed development. *PNAS* **114**:E9730–E9739. DOI: <https://doi.org/10.1073/pnas.1716758114>, PMID: 29078418
- Lindroth AM, Cao X, Jackson JP, Zilberman D, McCallum CM, Henikoff S, Jacobsen SE. 2001. Requirement of CHROMOMETHYLASE3 for maintenance of CpXpG methylation. *Science* **292**:2077–2080. DOI: <https://doi.org/10.1126/science.1059745>, PMID: 11349138
- Lippman Z, Gendrel AV, Black M, Vaughn MW, Dedhia N, McCombie WR, Lavine K, Mittal V, May B, Kasschau KD, Carrington JC, Doerge RW, Colot V, Martienssen R. 2004. Role of transposable elements in

- Heterochromatin and epigenetic control. *Nature* **430**:471–476. DOI: <https://doi.org/10.1038/nature02651>, PMID: 15269773
- Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR. 2008. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell* **133**:523–536. DOI: <https://doi.org/10.1016/j.cell.2008.03.029>, PMID: 18423832
- Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**:550. DOI: <https://doi.org/10.1186/s13059-014-0550-8>, PMID: 25516281
- Lutzmayr S, Enugutti B, Nodine MD. 2017. Novel small RNA spike-in oligonucleotides enable absolute normalization of small RNA-Seq data. *Scientific Reports* **7**:5913. DOI: <https://doi.org/10.1038/s41598-017-06174-3>, PMID: 28724941
- Maher KA. 2020. *Applications of Next-Generation Sequencing Strategies for the Identification and Characterization of Enhancers in Plants*. Emory University.
- Mathieu O, Jasencakova Z, Vaillant I, Gendrel AV, Colot V, Schubert I, Tourmente S. 2003. Changes in 5S rDNA chromatin organization and transcription during heterochromatin establishment in Arabidopsis. *The Plant Cell* **15**:2929–2939. DOI: <https://doi.org/10.1105/tpc.017467>, PMID: 14630972
- Mathieu O, Reinders J, Caikovski M, Smathajitt C, Paszkowski J. 2007. Transgenerational stability of the Arabidopsis epigenome is coordinated by CG methylation. *Cell* **130**:851–862. DOI: <https://doi.org/10.1016/j.cell.2007.07.007>, PMID: 17803908
- Mittelsten Scheid O, Afsar K, Paszkowski J. 1998. Release of epigenetic gene silencing by trans-acting mutations in Arabidopsis. *PNAS* **95**:632–637. DOI: <https://doi.org/10.1073/pnas.95.2.632>, PMID: 10939915
- Miura A, Nakamura M, Inagaki S, Kobayashi A, Saze H, Kakutani T. 2009. An Arabidopsis jmjC domain protein protects transcribed genes from DNA methylation at CHG sites. *The EMBO Journal* **28**:1078–1086. DOI: <https://doi.org/10.1038/emboj.2009.59>, PMID: 19262562
- Muyle A, Gaut BS. 2019. Loss of gene body methylation in *Eutrema salsugineum* is associated with reduced gene expression. *Molecular Biology and Evolution* **36**:155–158. DOI: <https://doi.org/10.1093/molbev/msy204>, PMID: 30398664
- Narsai R, Gouil Q, Secco D, Srivastava A, Karpievitch YV, Liew LC, Lister R, Lewsey MG, Whelan J. 2017. Extensive transcriptomic and epigenomic remodelling occurs during *Arabidopsis thaliana* germination. *Genome Biology* **18**:172. DOI: <https://doi.org/10.1186/s13059-017-1302-3>, PMID: 28911330
- Niederhuth CE, Bewick AJ, Ji L, Alabady MS, Kim KD, Li Q, Rohr NA, Rambani A, Burke JM, Udall JA, Egusi C, Schmutz J, Grimwood J, Jackson SA, Springer NM, Schmitz RJ. 2016. Widespread natural variation of DNA methylation within angiosperms. *Genome Biology* **17**:194. DOI: <https://doi.org/10.1186/s13059-016-1059-0>, PMID: 27671052
- Ning YQ, Liu N, Lan KK, Su YN, Li L, Chen S, He XJ. 2020. DREAM complex suppresses DNA methylation maintenance genes and precludes DNA hypermethylation. *Nature Plants* **6**:942–956. DOI: <https://doi.org/10.1038/s41477-020-0710-7>, PMID: 32661276
- Osakabe A, Lorkovic ZJ, Kobayashi W, Tachiwana H, Yelagandula R, Kurumizaka H, Berger F. 2018. Histone H2A variants confer specific properties to nucleosomes and impact on chromatin accessibility. *Nucleic Acids Research* **46**:7675–7685. DOI: <https://doi.org/10.1093/nar/gky540>, PMID: 29945241
- Osakabe A, Jamge B, Axelsson E, Montgomery SA, Akimcheva S, Kuehn AL, Pisupati R, Lorkovic ZJ, Yelagandula R, Kakutani T, Berger F. 2021. The chromatin remodeler DDM1 prevents transposon mobility through deposition of histone variant H2A.W. *Nature Cell Biology* **23**:391–400. DOI: <https://doi.org/10.1038/s41556-021-00658-1>, PMID: 33833428
- Papareddy RK, Paldi K, Paulraj S, Kao P, Lutzmayr S, Nodine MD. 2020. Chromatin regulates expression of small RNAs to help maintain transposon methylome homeostasis in Arabidopsis. *Genome Biology* **21**:251. DOI: <https://doi.org/10.1186/s13059-020-02163-4>, PMID: 32943088
- Papareddy R. 2021. RKP2021-CMT3. *Software Heritage*. sw:1:rev:89d7e2ea78af1969bb161640baed09296ed2485f. <https://archive.softwareheritage.org/sw:1:rev:89d7e2ea78af1969bb161640baed09296ed2485f>; <https://github.com/Gregor-Mendel-Institute/RKP2021-CMT3?visit=sw:1:sn:8e4b9655914607e3abe7868ed585c077572dcaf1&anchor=sw:1:rev:89d7e2ea78af1969bb161640baed09296ed2485f>
- Papareddy RK, Nodine MD. 2021. Plant epigenetics: propelling DNA methylation variation across the cell cycle. *Current Biology* **31**:R129–R131. DOI: <https://doi.org/10.1016/j.cub.2020.11.049>, PMID: 33561411
- Picard CL, Gehring M. 2017. Proximal methylation features associated with nonrandom changes in gene body methylation. *Genome Biology* **18**:73. DOI: <https://doi.org/10.1186/s13059-017-1206-2>, PMID: 28446217
- Picelli S, Faridani OR, Björklund AK, Winberg G, Sagasser S, Sandberg R. 2014. Full-length RNA-seq from single cells using Smart-seq2. *Nature Protocols* **9**:171–181. DOI: <https://doi.org/10.1038/nprot.2014.006>, PMID: 24385147
- Pignatta D, Erdmann RM, Scheer E, Picard CL, Bell GW, Gehring M. 2015. Natural epigenetic polymorphisms lead to intraspecific variation in Arabidopsis gene imprinting. *eLife* **e03198**. DOI: <https://doi.org/10.7554/eLife.08658>
- Plotnikova A, Kellner MJ, Schon MA, Mosiolek M, Nodine MD. 2019. MicroRNA dynamics and functions during Arabidopsis embryogenesis. *The Plant Cell* **31**:2929–2946. DOI: <https://doi.org/10.1105/tpc.19.00395>, PMID: 31562217
- Probst AV, Dunleavy E, Almouzni G. 2009. Epigenetic inheritance during the cell cycle. *Nature Reviews Molecular Cell Biology* **10**:192–206. DOI: <https://doi.org/10.1038/nrm2640>, PMID: 19234478

- Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**:841–842. DOI: <https://doi.org/10.1093/bioinformatics/btq033>, PMID: 20110278
- Ramírez F, Dündar F, Diehl S, Grüning BA, Manke T. 2014. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Research* **42**:W187–W191. DOI: <https://doi.org/10.1093/nar/gku365>, PMID: 24799436
- Rutowicz K, Lirski M, Mermaz B, Teano G, Schubert J, Mestiri I, Kroteń MA, Fabrice TN, Fritz S, Grob S, Ringli C, Cherkezyan L, Barneche F, Jerzmanowski A, Baroux C. 2019. Linker histones are fine-scale chromatin architects modulating developmental decisions in *Arabidopsis*. *Genome Biology* **20**:157. DOI: <https://doi.org/10.1186/s13059-019-1767-3>, PMID: 31391082
- Saze H, Mittelsten Scheid O, Paszkowski J. 2003. Maintenance of CpG methylation is essential for epigenetic inheritance during plant gametogenesis. *Nature Genetics* **34**:65–69. DOI: <https://doi.org/10.1038/ng1138>, PMID: 12669067
- Saze H, Shiraishi A, Miura A, Kakutani T. 2008. Control of genic DNA methylation by a jmjC domain-containing protein in *Arabidopsis thaliana*. *Science* **319**:462–465. DOI: <https://doi.org/10.1126/science.1150987>, PMID: 18218897
- Saze H, Kakutani T. 2007. Heritable epigenetic mutation of a transposon-flanked *Arabidopsis* gene due to lack of the chromatin-remodeling factor DDM1. *The EMBO Journal* **26**:3641–3652. DOI: <https://doi.org/10.1038/sj.emboj.7601788>, PMID: 17627280
- Saze H, Kakutani T. 2011. Differentiation of epigenetic modifications between transposons and genes. *Current Opinion in Plant Biology* **14**:81–87. DOI: <https://doi.org/10.1016/j.pbi.2010.08.017>, PMID: 20869294
- Schultz MD, He Y, Whitaker JW, Hariharan M, Mukamel EA, Leung D, Rajagopal N, Nery JR, Ulrich MA, Chen H, Lin S, Lin Y, Jung I, Schmitt AD, Selvaraj S, Ren B, Sejnowski TJ, Wang W, Ecker JR. 2015. Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature* **523**:212–216. DOI: <https://doi.org/10.1038/nature14465>, PMID: 26030523
- Shahzad Z, Moore JD, Zilberman D. 2021. Gene body methylation mediates epigenetic inheritance of plant traits. *bioRxiv*. DOI: <https://doi.org/10.1101/2021.03.15.435374>
- Shen JC, Rideout WM, Jones PA. 1992. High frequency mutagenesis by a DNA methyltransferase. *Cell* **71**:1073–1080. DOI: [https://doi.org/10.1016/S0092-8674\(05\)80057-1](https://doi.org/10.1016/S0092-8674(05)80057-1), PMID: 1473145
- Shen J, Xie K, Xiong L. 2010. Global expression profiling of rice microRNAs by one-tube stem-loop reverse transcription quantitative PCR revealed important roles of microRNAs in abiotic stress responses. *Molecular Genetics and Genomics* **284**:477–488. DOI: <https://doi.org/10.1007/s00438-010-0581-0>, PMID: 20941508
- Shepherd CT, Moran LA, Scott MP. 2009. Determination of transgene copy number by real-time quantitative PCR. *Methods in Molecular Biology* **526**:129–134. DOI: https://doi.org/10.1007/978-1-59745-494-0_11, PMID: 19378009
- Soppe WJ, Jasencakova Z, Houben A, Kakutani T, Meister A, Huang MS, Jacobsen SE, Schubert I, Franz PF. 2002. DNA methylation controls histone H3 lysine 9 methylation and heterochromatin assembly in *Arabidopsis*. *The EMBO Journal* **21**:6549–6559. DOI: <https://doi.org/10.1093/emboj/cdf657>, PMID: 12456661
- Stempor P, Ahringer J. 2016. *Wellcome Open Research*.
- Stroud H, Greenberg MV, Feng S, Bernatavichute YV, Jacobsen SE. 2013. Comprehensive analysis of silencing mutants reveals complex regulation of the *Arabidopsis* methylome. *Cell* **152**:352–364. DOI: <https://doi.org/10.1016/j.cell.2012.10.054>, PMID: 23313553
- Stroud H, Do T, Du J, Zhong X, Feng S, Johnson L, Patel DJ, Jacobsen SE. 2014. Non-CG methylation patterns shape the epigenetic landscape in *Arabidopsis*. *Nature Structural & Molecular Biology* **21**:64–72. DOI: <https://doi.org/10.1038/nsmb.2735>, PMID: 24336224
- Sved J, Bird A. 1990. The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. *PNAS* **87**:4692–4696. DOI: <https://doi.org/10.1073/pnas.87.12.4692>, PMID: 2352943
- Takuno S, Seymour DK, Gaut BS. 2017. The evolutionary dynamics of orthologs that shift in gene body methylation between *Arabidopsis* species. *Molecular Biology and Evolution* **34**:1479–1491. DOI: <https://doi.org/10.1093/molbev/msx099>, PMID: 28333248
- Takuno S, Gaut BS. 2013. Gene body methylation is conserved between plant orthologs and is of evolutionary consequence. *PNAS* **110**:1797–1802. DOI: <https://doi.org/10.1073/pnas.1215380110>, PMID: 23319627
- Toolkit P. 2019. *Broad Institute, GitHub Repository*. <http://broadinstitute.github.io/picard/>
- Tran RK, Henikoff JG, Zilberman D, Ditt RF, Jacobsen SE, Henikoff S. 2005. DNA methylation profiling identifies CG methylation clusters in *Arabidopsis* genes. *Current Biology* **15**:154–159. DOI: <https://doi.org/10.1016/j.cub.2005.01.008>, PMID: 15668172
- van Zanten M, Koini MA, Geyer R, Liu Y, Brambilla V, Bartels D, Koornneef M, Fransz P, Soppe WJJ. 2011. Seed maturation in *Arabidopsis thaliana* is characterized by nuclear size reduction and increased chromatin condensation. *PNAS* **108**:20219–20224. DOI: <https://doi.org/10.1073/pnas.1117726108>
- Wang L, Wang C, Liu X, Cheng J, Li S, Zhu J-K, Gong Z. 2019. Peroxisomal β^2 -oxidation regulates histone acetylation and DNA methylation in *Arabidopsis*. *PNAS* **116**:10576–10585. DOI: <https://doi.org/10.1073/pnas.1904143116>
- Wendte JM, Zhang Y, Ji L, Shi X, Hazarika RR, Shahryar Y, Johannes F, Schmitz RJ. 2019. Epimutations are associated with CHROMOMETHYLASE 3-induced de novo DNA methylation. *eLife* **8**:e47891. DOI: <https://doi.org/10.7554/eLife.47891>, PMID: 31356150
- Wierzbicki AT, Haag JR, Pikaard CS. 2008. Noncoding transcription by rna polymerase pol ivb/pol v mediates transcriptional silencing of overlapping and adjacent genes. *Cell* **135**:635–648. DOI: <https://doi.org/10.1016/j.cell.2008.09.035>

- Williams BP, Pignatta D, Henikoff S, Gehring M. 2015. Methylation-sensitive expression of a dna demethylase gene serves as an epigenetic rheostat. *PLOS Genetics* **11**:e1005142. DOI: <https://doi.org/10.1371/journal.pgen.1005142>
- Williams BP, Bechen LA, Pohlmann DA, Gehring M. 2021. Somatic DNA demethylation generates tissue-specific methylation states and impacts flowering time. *bioRxiv*. DOI: <https://doi.org/10.1101/2021.03.29.437569>
- Woo HR, Dittmer TA, Richards EJ. 2008. Three SRA-domain methylcytosine-binding proteins cooperate to maintain global cpG methylation and epigenetic silencing in *Arabidopsis*. *PLOS Genetics* **4**:e1000156. DOI: <https://doi.org/10.1371/journal.pgen.1000156>
- Xing H-L, Dong L, Wang Z-P, Zhang H-Y, Han C-Y, Liu B, Wang X-C, Chen Q-J. 2014. A CRISPR/Cas9 toolkit for multiplex genome editing in plants. *BMC Plant Biology* **14**:327. DOI: <https://doi.org/10.1186/s12870-014-0327-y>
- Yang L, Wang S, Tang L, Liu B, Ye W, Wang L, Wang Z, Zhou M, Chen B, Chen B-C. 2014. Universal Stem-Loop Primer Method for Screening and Quantification of MicroRNA. *PLOS ONE* **9**:e115293. DOI: <https://doi.org/10.1371/journal.pone.0115293>
- Yelagandula R, Stroud H, Holec S, Zhou K, Feng S, Zhong X, Muthurajan UM, Nie X, Kawashima T, Groth M, Luger K, Jacobsen SE, Berger F. 2014. The histone variant H2A.W defines heterochromatin and promotes chromatin condensation in *Arabidopsis*. *Cell* **158**:98–109. DOI: <https://doi.org/10.1016/j.cell.2014.06.006>, PMID: 24995981
- Zemach A, Kim M, Hsieh P-H, Coleman-Derr D, Eshed-Williams L, Thao K, Harmer SL, Zilberman D. 2013. The *Arabidopsis* Nucleosome Remodeler DDM1 Allows DNA Methyltransferases to Access H1-Containing Heterochromatin. *Cell* **153**:193–205. DOI: <https://doi.org/10.1016/j.cell.2013.02.033>
- Zhai J, Bischof S, Wang H, Feng S, Lee TF, Teng C, Chen X, Park SY, Liu L, Gallego-Bartolome J, Liu W, Henderson IR, Meyers BC, Ausin I, Jacobsen SE. 2015. A one precursor one siRNA model for pol IV-Dependent siRNA biogenesis. *Cell* **163**:445–455. DOI: <https://doi.org/10.1016/j.cell.2015.09.032>, PMID: 26451488
- Zhang X, Yazaki J, Sundaresan A, Cokus S, Chan SW-L, Chen H, Henderson IR, Shinn P, Pellegrini M, Jacobsen SE, Ecker JR. 2006. Genome-wide high-resolution mapping and functional analysis of dna methylation in *Arabidopsis*. *Cell* **126**:1189–1201. DOI: <https://doi.org/10.1016/j.cell.2006.08.003>
- Zhang Y, Wendte JM, Ji L, Schmitz RJ. 2020. Natural variation in DNA methylation homeostasis and the emergence of epialleles. *PNAS* **117**:4874–4884. DOI: <https://doi.org/10.1073/pnas.1918172117>
- Zhong Z, Feng S, Duttke SH, Potok ME, Zhang Y, Gallego-Bartolome J, Liu W, Jacobsen SE. 2021. DNA methylation-linked chromatin accessibility affects genomic architecture in *Arabidopsis*. *PNAS* **118**:e2023347118. DOI: <https://doi.org/10.1073/pnas.2023347118>
- Zilberman D. 2003. ARGONAUTE4 control of Locus-Specific siRNA accumulation and DNA and histone methylation. *Science* **299**:716–719. DOI: <https://doi.org/10.1126/science.1079695>
- Zilberman D, Gehring M, Tran RK, Ballinger T, Henikoff S. 2007. Genome-wide analysis of *Arabidopsis thaliana* DNA methylation uncovers an interdependence between methylation and transcription. *Nature Genetics* **39**:61–69. DOI: <https://doi.org/10.1038/ng1929>
- Zilberman D. 2017. An evolutionary case for functional gene body methylation in plants and animals. *Genome Biology* **18**:87. DOI: <https://doi.org/10.1186/s13059-017-1230-2>, PMID: 28486944

4 GSNOR Contributes to Demethylation and Expression of Transposable Elements and Stress-Responsive Genes

status	<i>published</i> Antioxidants 10, 7, 1128 (2021) doi:10.3390/antiox10071128
authors	Eva-Esther Rudolf (E.E.R), Patrick Hüther (P.H.), Ignasi Forné (I.F.), Elisabeth Georgii (E.G.), Yongtao Han (Y.H.), Rüdiger Hell (R.H.), Markus Wirtz (M.W.), Axel Imhof (A.I.), Claude Becker (C.B.), Jörg Durner (J.D.) and Christian Lindermayr (C.L.)
contributions	Conceptualization, C.L.; formal analysis, E.E.R., P.H., I.F., E.G., and M.W.; investigation, E.E.R., P.H., I.F., Y.H.,and M.W.; methodology, E.E.R., P.H.,and M.W.; supervision, C.L.; Writing — original draft, E.E.R.; Writing — review and editing, R.H., A.I., C.B., J.D.,and C.L.



Article

GSNOR Contributes to Demethylation and Expression of Transposable Elements and Stress-Responsive Genes

Eva Esther Rudolf ¹, Patrick Hühner ^{2,3}, Ignasi Forné ⁴, Elisabeth Georgii ¹, Yongtao Han ¹, Rüdiger Hell ⁵, Markus Wirtz ⁵, Axel Imhof ⁴, Claude Becker ^{2,3}, Jörg Durner ^{1,6} and Christian Lindermayr ^{1,*}

- ¹ Institute of Biochemical Plant Pathology, Helmholtz Zentrum München, German Research Center for Environmental Health, 85764 Oberschleißheim, Germany; eva.rudolf@mailbox.org (E.R.); elisabeth.georgii@helmholtz-muenchen.de (E.G.); hanyongtao515@163.com (Y.H.); durner@helmholtz-muenchen.de (J.D.)
- ² Gregor Mendel Institute of Molecular Plant Biology, Austrian Academy Sciences, Vienna BioCenter (VBC), 1030 Wien, Austria; patrick.huether@gmi.oeaw.ac.at (P.H.); claude.becker@biologie.uni-muenchen.de (C.B.)
- ³ Genetics, LMU Biocenter, Ludwig-Maximilians-Universität München, 82152 Martinsried, Germany
- ⁴ Protein Analysis Unit, Ludwig-Maximilians-Universität München, 82152 Martinsried, Germany; ignasi.forne@lrz.uni-muenchen.de (I.F.); imhof@lmu.de (A.I.)
- ⁵ Centre for Organismal Studies, Ruprecht-Karls-Universität Heidelberg, 69120 Heidelberg, Germany; ruediger.hell@cos.uni-heidelberg.de (R.H.); markus.wirtz@cos.uni-heidelberg.de (M.W.)
- ⁶ Chair of Biochemical Plant Pathology, Technische Universität München, 85354 Freising, Germany
- * Correspondence: lindermayr@helmholtz-muenchen.de; Tel.: +49-89-31872285



Citation: Rudolf, E.E.; Hühner, P.; Forné, I.; Georgii, E.; Han, Y.; Hell, R.; Wirtz, M.; Imhof, A.; Becker, C.; Durner, J.; et al. GSNOR Contributes to Demethylation and Expression of Transposable Elements and Stress-Responsive Genes. *Antioxidants* **2021**, *10*, 1128. <https://doi.org/10.3390/antiox10071128>

Academic Editors: María Cruz González and Antonio Jesús Serrato

Received: 9 June 2021
Accepted: 9 July 2021
Published: 15 July 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Abstract: In the past, reactive nitrogen species (RNS) were supposed to be stress-induced by-products of disturbed metabolism that cause oxidative damage to biomolecules. However, emerging evidence demonstrates a substantial role of RNS as endogenous signals in eukaryotes. In plants, S-nitrosoglutathione (GSNO) is the dominant RNS and serves as the •NO donor for S-nitrosation of diverse effector proteins. Remarkably, the endogenous GSNO level is tightly controlled by S-nitrosoglutathione reductase (GSNOR) that irreversibly inactivates the glutathione-bound NO to ammonium. Exogenous feeding of diverse RNS, including GSNO, affected chromatin accessibility and transcription of stress-related genes, but the triggering function of RNS on these regulatory processes remained elusive. Here, we show that GSNO reductase-deficient plants (*gsnor1-3*) accumulate S-adenosylmethionine (SAM), the principal methyl donor for methylation of DNA and histones. This SAM accumulation triggered a substantial increase in the methylation index (MI = [SAM]/[S-adenosylhomocysteine]), indicating the transmethylation activity and histone methylation status in higher eukaryotes. Indeed, a mass spectrometry-based global histone profiling approach demonstrated a significant global increase in H3K9me2, which was independently verified by immunological detection using a selective antibody. Since H3K9me2-modified regions tightly correlate with methylated DNA regions, we also determined the DNA methylation status of *gsnor1-3* plants by whole-genome bisulfite sequencing. DNA methylation in the CG, CHG, and CHH contexts in *gsnor1-3* was significantly enhanced compared to the wild type. We propose that GSNOR1 activity affects chromatin accessibility by controlling the transmethylation activity (MI) required for maintaining DNA methylation and the level of the repressive chromatin mark H3K9me2.

Keywords: nitric oxide; S-nitrosoglutathione; S-nitrosoglutathione reductase; metaboloepigenetic; S-adenosylhomocysteine; DNA methylation; histone methylation

1. Introduction

Nitric oxide (•NO) is a ubiquitous signaling molecule with pleiotropic functions throughout the lifespan of plants. Indeed, •NO is involved in the regulation of growth and development processes including seed dormancy [1], seed germination [2,3], root growth [4,5], hypocotyl elongation [6], stomatal closure [7], gravitropism [8], flowering [9,10], pollen tube growth [11], fruit ripening and senescence [12], biotic and abiotic stress responses [13–15],

and iron homeostasis [16]. In plants, •NO is endogenously produced in different cellular compartments, including the cytosol, peroxisomes, mitochondria, and chloroplasts [17], under both physiological and stress conditions [18].

Although •NO biosynthesis has not been described in the nucleus, it is possibly transferred into the nucleus by passive diffusion, through S-nitrosated proteins or S-nitrosated low-molecular weight thiols, such as S-nitrosoglutathione (GSNO) (discussed in [19]).

•NO and reactive nitrogen species (RNS) exert their biological function through post-translational modifications (PTMs), including tyrosine nitration, metal nitrosylation, and S-nitrosation. In general, those •NO-mediated PTMs have profound effects on the function of target proteins by regulating their activities, subcellular localization, structure, or interaction with biomolecules [20–22]. Protein S-nitrosation is the most important •NO-mediated PTM [14]. Proteome-wide studies identified putatively S-nitrosated proteins involved in numerous aspects of plant biology, such as plant immune response, the antioxidant system, metabolic processes, and transcription factors. Consequently, •NO regulates diverse physiological processes by altering gene expression [23–25], metabolite levels [26], and/or phytohormone signaling [27,28].

•NO is a short-lived free-radical, whose function is restricted to its local microenvironment. In contrast, GSNO is a more stable redox form of •NO [29,30] regarded as an intracellular mobile •NO reservoir [31], which can release •NO in the presence of metal ions, such as copper and iron, or reductants, such as ascorbate or GSH [32]. Moreover, GSNO can transfer its •NO moiety directly to cysteine thiol groups of other proteins (trans-nitrosation) [22,29]. GSNO levels are controlled by the activity of S-nitrosoglutathione reductase (GSNOR). GSNOR is an evolutionarily conserved enzyme that catalyzes the NADH-dependent reduction of GSNO to oxidized GSH (GSSG) and ammonia in the presence of GSH [29,30]. The *Arabidopsis* genome encodes a single-copy GSNOR gene. Loss of GSNOR1 leads to elevated levels of •NO, nitrate, nitrite, and S-nitrosothiols (RSNOs), and a proteome-wide increased S-nitrosation [33–39]. Hence, GSNOR1 is considered to control the intracellular levels of both GSNO and, indirectly, protein SNOs [34]. GSNOR1 deficiency causes pleiotropic plant growth and development defects, impaired plant disease responses, heat sensitivity, and resistance to cell death [29–31], suggesting a regulatory role of GSNOR in these processes [29–31].

A number of nuclear proteins that undergo S-nitrosation have been identified [40], suggesting a regulatory function of GSNO/•NO in nuclear events/processes. Apart from the transcriptional or post-translational control of transcription factors [41], several lines of evidence demonstrate that GSNO/•NO regulates gene expression also via modulation of the chromatin structure and/or DNA accessibility [23]. In general, the distinct chromatin states that modulate access to DNA for transcription are regulated by multiple epigenetic mechanisms, including DNA methylation, covalent modifications of core histones such as methylation and acetylation, ATP-dependent chromatin remodeling, placement of histone variants, non-coding RNAs, and metaboloeipigenetic effects [42,43]. In *Arabidopsis*, GSNO treatment induced histone hyperacetylation at genes related to stress response by inhibiting histone deacetylase activity [44]. Further, the plant-specific histone deacetylases HDT1/2 regulating the expression of *GIBBERELLIN 2-OXIDASE2* by histone acetylation [45] were identified as targets for S-nitrosation [40]. DNA hypomethylation concomitant with transcriptional activation of transposable elements (TEs) was observed in rice upon treatment with 0.1–1 mM of the •NO donor sodium nitroprusside [46].

From bacteria to humans, methylation is sensitive to the cellular metabolic status [47]. Both the methylation cycle and the tricarboxylic acid cycle provide substrates for enzymes involved in DNA and histone methylation. Indeed, methylation is directly linked to intermediary metabolism with S-adenosylmethionine (SAM) acting as the main methyl donor for transmethylation reactions catalyzed by methyltransferases, which methylate DNA, RNA, lipids, histones, and cellular metabolites [48]. Each transmethylation reaction consumes SAM and releases the by-product S-adenosylhomocysteine (SAH). SAH, a competi-

tive inhibitor of methyltransferases, is then recycled to homocysteine (Hcys) and adenosine by S-adenosylhomocysteine hydrolase (SAHH). The equilibrium of this reversible reaction favoring SAH is driven towards hydrolysis of SAH due to removal of its products by downstream enzymes. Methionine synthase (MS) catalyzes the methylation of Hcys to methionine using methyl-tetrahydrofolate (CH₃-THF). Then, S-adenosylmethionine synthetase (SAMS) catalyzes the adenylation of methionine to SAM to close the methylation cycle [48,49]. The recycling mechanism is crucial for maintaining an adequate methylation index (MI; SAM/SAH ratio), which is regarded as an indicator of the cellular methylation state.

Numerous studies reported that SAM and SAH levels regulate DNA and histone methylation [42,50,51]. The *Arabidopsis* genome encodes two *SAHH* isoforms; however, *SAHH1* is assumed to play a predominant role in maintaining TGS and DNA methylation at numerous targets compared to *SAHH2* [52,53]. *Arabidopsis sahh1* knock-down mutants (*sahh1-kd*; knockout is zygotic lethal [52]) possessed a decreased MI [52,54], as well as decreased DNA and H3K9me2 methylation, concomitant with the release of transcriptional silencing at transgene reporters [52,53], repetitive DNA sequences such as ribosomal DNA and 180 bps repeats [52–54], and transposons [55,56]. Similarly, the expression of antisense RNA of *SAHH* in tobacco plants resulted in a loss of DNA methylation in repetitive elements [57]. Other studies employed a selective reversible inhibitor of SAHH, namely, dihydroxypropyladenine (DHPA). In tobacco, DHPA caused accumulation of SAH and DNA hypomethylation [58–60]. In *Arabidopsis*, the application of DHPA reduced levels of DNA and histone methylation at endogenous repeats [53]. Moreover, *SAMS4* is an important epigenetic regulator in *Arabidopsis*. Mutations in *SAMS4* caused decreased SAM levels, CHG/CHH and H3K9me2 hypomethylation, and activation of TEs [61]. Similarly, *MS1* mutation resulted in a decreased MI, and decreased DNA and H3K9me2 hypomethylation [50]. Accordingly, overexpression of *MS1* is accompanied by a genome-wide global increase in DNA methylation in *Arabidopsis* [62].

Here, we report that the GSNOR1 function is required for SAM homeostasis, and, hence, for balancing the methylation index (ratio of SAM/SAH). Consequently, loss of GSNOR1 activity affects transmethylation reactions. Nano-liquid chromatography mass spectrometry (LC-MS) profiling of histone modifications demonstrated a significant global increase in the repressive H3K9me2 mark in *gsnor1-3*. Whole-genome bisulfite sequencing and transcriptome analyses revealed enhanced DNA methylation and reduced expression of TEs and stress-responsive genes in *gsnor1-3*, in comparison to the wild type. Our data suggest that the GSNOR1 function is required to reduce the level of the repressive chromatin mark H3K9me2, which is associated with the silencing of repeats and TEs. This function might be link to the activation of stress response genes.

2. Materials and Methods

2.1. Plant Material and Cultivation

A. thaliana ecotype Columbia-0 (Col-0; wt) purchased from the Nottingham Arabidopsis Stock Center (NASC), *gsnor1-3* obtained from GABI-Kat (also named *hot5-2*, GABI-Kat 315D11), *sahh1* purchased from NASC (SALK 068487), and the *A. thaliana* Col-0 *TS-GUS* (possesses a transcriptionally silent (TS), highly repetitive β-glucuronidase (GUS) transgene; L5, 6b5) line kindly provided by Hervé Vaucheret were used in this study and were previously described [34,35,53,54,56,63,64]. The *A. thaliana* Col-0 *TS-GUS* (L5, 6b5) line [64] was crossed with the mutants *sahh1* and *gsnor1-3*. The segregating F2 plants were genotyped, and seeds from lines homozygous for the *TS-GUS* locus and the mutation were used for further analysis. Oligonucleotides are listed in Supplemental Table S1.

Arabidopsis plants were grown on soil mixed with silica sand in a ratio of 4:1 in 4-well plant pots placed in a tray. Before sowing, soil was wetted with water supplemented with 0.15% (*v/v*) Neudorff Neudomück®. After stratification for two days at 4 °C in the dark, plants were cultivated for four weeks in a climate chamber at 65–68% relative humidity under long-day conditions (14 h light/10 h dark cycle, 20 °C day/18 °C night regime,

70 $\mu\text{mol m}^{-2} \text{s}^{-1}$ photon flux density). Four-week-old rosette leaves were harvested 5 h after the day-time start and flash frozen in liquid nitrogen.

For liquid culture experiments, *A. thaliana* seeds were surface sterilized by soaking in 70% (v/v) ethanol for 1 min and then in 50% (v/v) household bleach for 10 min followed by five washes with sterile ddH₂O. Seeds were suspended in sterile water and stratified for 2 days at 4 °C in the dark. Seedlings were cultivated in six-well plates containing 5 mL of 1x Murashige and Skoog (MS) medium [65] adjusted to pH 5.7 with potassium hydroxide and supplemented with 1% sucrose and 0.5 g L⁻¹ MES. Plantlets were germinated and grown for twelve days in liquid media supplemented with 200 μM DHPA or water (mock) under short-day conditions (10 h light/14 h dark cycle, 16 °C day/ 20 °C night regime, relative humidity 80% day/65% night, 100 $\mu\text{mol m}^{-2} \text{s}^{-1}$ photon flux density) on a shaker (100 rpm). Media including drugs were exchanged every day at the night-time start.

2.2. Quantification of MTA, SAM, SAH, Cys GSH, and Hcys

Samples (0.1 g) were ground in liquid nitrogen to a fine powder and extracted with 0.1 M HCl (0.1 mL) by vortexing at 4 °C for 15 min. The resulting homogenates were centrifuged for 10 min at 4 °C and 16,400 g to remove cell debris. Adenosines were derivatized with chloro-acetaldehyde as previously described [66]. The metabolites were separated by reversed-phase chromatography on an Acquity HSS T3 column (100 mm × 2.1 mm, 1.7 μm , Waters) equilibrated in buffer A (5.7 mM TBAS, 30.5 mM KH₂PO₄ pH 5.8) by applying the following gradient: 0.6 mL 1% B, 1.9 mL 8% B, 1.9 mL 14% B, 5.7 mL 50% B. Buffer B was a mix of 34% buffer A and 66% acetonitrile. The fluorescent 1,N⁶-etheno-derivatives of MTA, SAM, and SAH were quantified with an Acquity FLR detector (Waters, excitation: 280 nm, emission: 410 nm) connected to an H-class UPLC system. The thiols (Cys, GSH, and Hcys) were labeled with monobromobimane, and the resulting fluorescent thiol-bimane derivatives were separated by reversed-phase chromatography according to (Dong et al., 2017).

2.3. Whole-Genome Bisulfite Sequencing and Data Analysis

WGBS library preparation and sequencing. WGBS was performed from snap-frozen 4-week-old rosette leaves grown under long-day conditions and harvested 5 h after the day-time start (total 1.5 g) for each genotype. Two biological replicates were analyzed for each genotype. gDNA was extracted from leaf samples (aliquot 150 mg, ground in liquid nitrogen) with DNeasy® Plant Mini Kit and sheared to 350 bps. WGBS DNA libraries were generated using the Illumina® TruSeq® Nano Kit, and bisulfite treatment was conducted with the EpiTect® Plus Bisulfite Kit. Briefly, the fragmented DNAs were end repaired, adenine bases were added to the 3' end (A-tailing) of the DNA fragments, and methylated adapters were ligated to the DNA fragment. Next, the DNA fragments were size selected before sodium bisulfite treatment and PCR amplification (KAPA HiFi HS Uracil+ R from Roche Cat.No:795905001). Libraries were sequenced with 125 bp paired-end reads on a HiSeq 2500 instrument.

Processing and alignment of bisulfite-converted reads. For read mapping, the nf-core methylseq pipeline available at <https://github.com/nf-core/methylseq> (accessed on 12 February 2018 (<https://doi.org/10.5281/zenodo.1343417>; accessed on 9 May 2021)) was used. In short, raw sequencing reads were quality controlled (FastQC v0.11.5), and sequencing adapters were trimmed off (Trim Galore v0.4.1). Reads were aligned to the TAIR9 Reference genome with Bismark (version v0.17.0) [67] using the Bowtie2 aligner [68]. After deduplication (picardtool MarkDuplicates v2.8.0), methylated Cs were extracted from aligned reads with MethylExtract (v1.9.1). Bisulfite conversion efficiency was calculated from the proportion of unconverted Cs in the chloroplast genome.

Post-alignment Analysis. Methylation calling information of each individual cytosine was tabulated and subjected to post-alignment analysis with the MethylScore pipeline. Briefly, identification of differentially methylated positions was performed according to Becker et al. [69]. Identification of methylated regions (MRs) and differentially methylated

regions (DMRs) was conducted by an adaption of a hidden Markov model-based approach, as previously described [70], which identifies regions of dense methylation that are then tested for differential methylation [71]. The DMRs were identified by pairwise comparison of WGBS profiles (*gsnor1-3* vs. Col-0/wt; *sahh1* vs. Col-0/wt).

Annotation—mapping to genomic elements. For annotation of genomic elements, the TAIR10 reference annotation was used. MRs and DMRs were assigned to annotated elements (CDS, intron, 5' UTR, 3' UTR, transposon, 2kb upstream, 2kb downstream, as-lncRNA, lncRNA, miRNA, pri-miRNA, ncRNA, snoRNA, tRNA, pseudogene). Genes with at least one DMR in the gene body, at 3kb up- or downstream of flanking regions, were considered as differentially methylated genes (DMGs). Further, TEs with at least one DMR were identified.

2.4. RNA Sequencing

RNA-seq was performed from snap-frozen 4-week-old rosette leaves grown under long-day conditions and harvested 5 h after the day-time start (total 1.5 g) for each genotype. Four replicates were analyzed for each genotype. RNA was extracted from 4-week-old rosette leaves using the innuPREP PLANT RNA Kit. Sequencing libraries were generated from Poly(A)-enriched RNA using the NEBNext[®] Ultra[™] II Directional RNA Library Prep kit (New England Biolabs) according to the manufacturer's instructions and sequenced on a HiSeqV4 instrument (Illumina) as 100 bp single-end reads. Reads were mapped to the TAIR10 reference of *Arabidopsis thaliana* annotated genes (www.arabidopsis.org; accessed on 24 December 2019) using STAR (v2.5.2a) [72]. Read quantifications were generated using Kallisto (v0.43.1) [73]. Differential expression analysis was performed using the DESeq2 package (v1.18.1) in R [74]. Gene annotation was performed using the following sources: UniProtKB, Swiss-Prot, TrEMBL, and TAIR.

2.5. Acid Extraction of Histones

Nuclei from 4-week-old rosette leaves were purified as described previously [75], with minor modifications. Two grams of plant tissue was grinded to a fine powder in liquid nitrogen, homogenized in two volumes of lysis buffer (20 mM Tris-HCl pH 7.4, 25% (v/v) glycerol, 20 mM KCl, 2 mM EDTA, 2.5 mM MgCl₂, 250 mM sucrose) supplemented with protease inhibitor, and incubated for 10 min on ice with intermittent vortexing. The homogenate was successively filtered through miracloth and a 160 µm nylon mesh. The flow-through was centrifuged at 1500 g for 10 min at 4 °C, and the pellet was washed four times with 4 mL of nuclear resuspension buffer (20 mM Tris-HCl pH 7.4, 25% (v/v) glycerol, 2.5 mM MgCl₂, 0.2% (v/v) Triton[®] X-100). The final pellet was resuspended in 700 µL of 0.2 M sulfuric acid to extract histones and other acid-soluble proteins overnight using an overhead shaker. The extract was then centrifuged at 16,000 × g for 20 min at 4 °C. The supernatant containing core histones was transferred to a new tube, and proteins were precipitated with 26% (v/v) trichloroacetic acid for 3 h on ice. After four washes with ice-cold acetone, the histone pellet was resuspended in 60 µL of 2× sample buffer (4% (w/v) SDS, 20% (v/v) glycerol, 10% (v/v) 2-mercaptoethanol, 0.004% (w/v) bromo-phenol blue, and 0.125 M Tris-HCl, pH 6.8) at 1500 rpm for 20 min at RT using a thermoshaker and then stored at −20 °C overnight. If necessary, pH was adjusted with 1 µL of 1 M Tris-HCl pH 8.0. Histones were separated by 12% SDS-PAGE (loading volume: 15–25 µL per lane) and stained with Coomassie[®] Blue for LC-MS/MS analysis or immunoblotted.

2.6. Quantification of Histone Methylation and Acetylation by LC-MS/MS

After electrophoreses, bands corresponding to histones H3 and H4 were excised from the gels. Calf histones were run as a size marker. Destaining, d3-acylation, digestion, and peptide desalting were performed as described before [76], with minor modifications. Desalted histone peptides in 0.1% TFA were injected in an RSLCnano system (Thermo Fisher Scientific) and separated in a 15 cm analytical column (75 µm ID home-packed with ReproSil-Pur C18-AQ 2.4 µm from Dr. Maisch), with a 50 min gradient from 4 to 40%

acetonitrile in 0.1% formic acid at a 300 nL/min flowrate. The effluent from the HPLC was electrosprayed into a Q Exactive HF mass spectrometer (Thermo Fisher Scientific). The MS instrument was programmed to target several ions as previously described [76], except for the MS3 fragmentation. Survey full-scan MS spectra (from m/z 270–730) were acquired with resolution $R = 60,000$ at m/z 400 (AGC target of 3×10^6). Targeted ions were isolated with an isolation window of 0.7 m/z to a target value of 2×10^5 and fragmented at 27% normalized collision energy. Typical mass spectrometric conditions were: spray voltage, 1.5 kV; no sheath and auxiliary gas flow; heated capillary temperature, 250 °C.

Peptide and fragment masses of histone H3 methylation and acetylation marks were calculated in GPMW [77]. The histone PTMs were quantified based on the area of the peak from the extracted ion chromatogram, and the Xcalibur™ software v2.2 SP1 (Quan Browser, Thermo Fisher Scientific) was used. The theoretical mass to charge ratio was calculated with GPMW 5.02 for each peptide. Further parameters were peak detection: Genesis; trace: mass range; mass tolerance: 20 ppm; mass precision: 4 decimals; S/N threshold: 0.5. After peak integration, data were exported to Excel, and the relative abundance was calculated as previously described [76].

2.7. Accession Numbers

Assigned accession numbers for the genes used in this work from The Arabidopsis Information Resource (www.arabidopsis.org accessed on 24 December 2019) are as follows: AT4G13940 (*SAHH1*), AT5G43940 (*GSNOR1*). RNA-seq and WGBS data were stored in the European Nucleotide Archive (ENA) under the ENA accession number PRJEB43942.

3. Results

3.1. GSNOR and SAHH1 Are Involved in Regulating Metabolite Levels of the Methylation Cycle

The function of GSNOR bioactivity in regulating metabolite levels of the methylation cycle was investigated in vivo using a GSNOR-deficient line, namely, *gsnor1-3* [34,35]. The GSNOR1-deficient mutant (*gsnor1-3*; also named *hot5-2*) is an important tool for functional analysis of GSNO under physiological conditions. Since in this mutant, the enzymatic degradation of GSNO is abolished, the observed phenotypic and molecular effects directly reflect the functions of SNOs in vivo [29,31]. Besides the Col-0 wild type (wt), the *sahh1* knock-down mutant [54,56,63] was used as a control plant (Supplemental Figure S1A). Both, GSNOR activity and the total RSNO content were analyzed in *gsnor1-3* and *sahh1* mutants under basal conditions. GSNOR activity significantly decreased to 10% in *gsnor1-3* relative to wt plants (Supplemental Figure S1B). The decreased GSNOR activity in the *gsnor1-3* mutant was accompanied by an almost 2-fold increase in RSNO levels (Supplemental Figure S1C). These results are in accordance with previous studies [34,35,39]. Neither the GSNOR activity nor the RSNO content is significantly different between *sahh1* and wt (Supplemental Figure S1B,C).

To analyze whether GSNOR1 and SAHH1 functions are required for intact processing of the methylation cycle and connected pathways, we determined the steady-state levels of SAM, SAH, and Hcys (methylation cycle), cysteine (Cys) and glutathione (GSH) (glutathione biosynthesis), and 5'-methylthioadenosin (MTA; by-product of polyamine, ethylene, and nicotianamine biosynthesis) in wt, *sahh1*, and *gsnor1-3* (Figure 1A–B,D–F). Moreover, we calculated the SAM/SAH ratio, also named the methylation index (MI), which is regarded as an indicator of the cellular methylation state (Figure 1C). The major methyl group donor SAM was significantly elevated in *sahh1* and *gsnor1-3* plants by 61% and 43%, respectively (Figure 1A). The *sahh1* mutant also showed a 2-fold increase in the SAH level, resulting in an overall decrease in the MI by 14% compared to wt (Figure 1B,C). This result is concordant with previous studies [52,54]. Since the SAH level of *gsnor1-3* was similar to that of wt, the resulting SAM/SAH ratio was significantly increased by 47% (Figure 1C). Hcys levels were below the detection limit in all three lines (data not shown), whereas the levels of MTA, Cys, and GSH were significantly increased in *gsnor1-3* and *sahh1* in comparison to wt (Figure 1D–F). In conclusion, target metabolomic analysis in wt, *sahh1*,

and *gsnor1-3* revealed alterations in the levels of SAM, SAH, MTA, Cys, and GSH, acting as precursors of substrates, cofactors, or inhibitors in epigenetic methylation processes. These results suggest that SAHH1 and GSNOR1 functions might be linked to histone and DNA methylation.

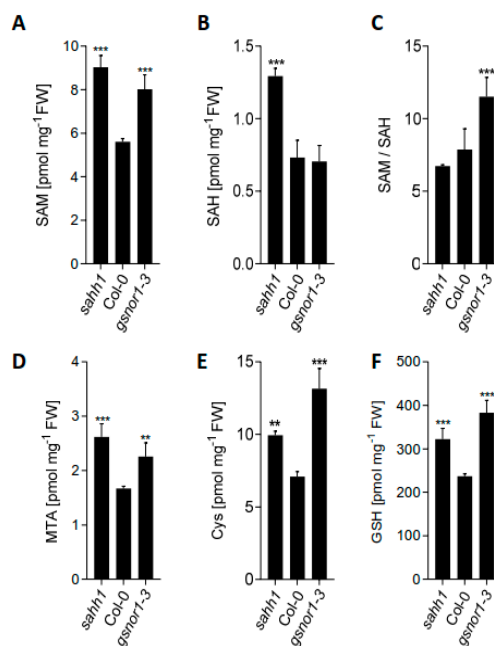


Figure 1. Mutations in *GSNOR1* and *SAHH1* result in an impaired methylation cycle. Analysis of steady-state levels of (A) SAM, (B) SAH, (C) SAM/SAH, (D) MTA, (E) Cys, and (F) GSH in 4-week-old rosette leaves grown under long-day conditions and harvested 5 h after the day-time start ($n = 5$). Values are normalized against total fresh weight and represent the mean $\pm$ SD. Grubb's outlier test ($\alpha = 0.05$) was performed. ** ($p < 0.01$) and *** ($p < 0.001$) represent significant differences between wt and mutants (ANOVA with Dunnett's multiple comparisons test). Statistical analysis was performed with GraphPad Prism version 7.05.

Interestingly, SAHH1 was identified as S-nitrosated under basal and stress conditions in proteome-wide studies [33,78–81] and in *gsnor1-3* seedlings [33], and several groups demonstrated that tyrosine nitration and S-nitrosation strongly inhibit SAHH1 activity in vitro [82]. We confirmed that recombinant SAHH1 can be S-nitrosated and reversibly inhibited by GSNO (Supplemental Figure S2A,B). SAHH1 is also inhibited by the sulfhydryl-modifying agent N-ethylmaleimide (NEM), confirming that cysteine residues are important for its activity (Supplemental Figure S2B). However, although *gsnor1-3* has an enhanced level of RSNOs, in vivo S-nitrosation of SAHH1 could not be detected (Supplemental Figure S2C). This, together with the fact that SAH (and Hcys) levels are unchanged in *gsnor1-3*, in comparison to wt, suggests that loss of the GSNOR function is not linked to inhibition of SAHH1 under the analyzed conditions.

3.2. Loss of GSNOR1 and SAHH1 Functions Results in Altered Histone Methylation Levels

To investigate the consequence of the altered metabolite levels and the MI in *gsnor1-3* and *sahh1* on histone modification, 4-week-old rosette leaves were analyzed by LC-MS [76].

The analysis of histone H3 revealed that the H3K9me2 level significantly increased by 23% and significantly decreased by 34% in *gsnor1-3* and *sahh1*, respectively, relative to wt (Table 1).

Table 1. Histone H3K9me2 methylation level is altered in *gsnor1-3* and *sahh1*.

Motif	Sequence of Peptide	Mean % Abundance $\pm$ SD		
		<i>sahh1</i>	Col-0/wt	<i>gsnor1-3</i>
H3.K4_noPTM	TKQTAR	42.81 $\pm$ 1.81	43.64 $\pm$ 3.12	42.70 $\pm$ 3.94
H3.K4me1	TKme1QTAR	56.71 $\pm$ 1.75	55.71 $\pm$ 2.56	56.43 $\pm$ 3.63
H3.K4me2	TKme2QTAR	0.21 $\pm$ 0.05	0.29 $\pm$ 0.24	0.40 $\pm$ 0.16
H3.K4me3	TKme3QTAR	0.27 $\pm$ 0.08	0.36 $\pm$ 0.3	0.47 $\pm$ 0.19
H3.K9_K14_noPTM	KSTGGKAPR	42.62 $\pm$ 0.65 **	39.68 $\pm$ 0.26	37.09 $\pm$ 0.78 **
H3.K9ac	KacSTGGKAPR	1.97 $\pm$ 0.31	1.55 $\pm$ 0.18	1.58 $\pm$ 0.10
H3.K14ac	KSTGGKacAPR	27.07 $\pm$ 0.79	27.02 $\pm$ 0.19	27.46 $\pm$ 0.56
H3.K9ac_K14ac	KacSTGGKacAPR	3.02 $\pm$ 0.12	3.13 $\pm$ 0.13	2.85 $\pm$ 0.20
H3.K9me1_K14ac	Kme1STGGKacAPR	1.17 $\pm$ 0.06	1.13 $\pm$ 0.08	1.24 $\pm$ 0.14
H3.K9me2_K14ac	Kme2STGGKacAPR	0.16 $\pm$ 0.03	0.16 $\pm$ 0.06	0.22 $\pm$ 0.06
H3.K9me3_K14ac	Kme3STGGKacAPR	0.02 $\pm$ 0.00	0.03 $\pm$ 0.02	0.03 $\pm$ 0.01
H3.K9me1	Kme1STGGKAPR	20.01 $\pm$ 0.34	21.31 $\pm$ 0.61	22.14 $\pm$ 0.50
H3.K9me2	Kme2STGGKAPR	3.88 $\pm$ 0.09 ***	5.83 $\pm$ 0.38	7.17 $\pm$ 0.41 **
H3.K9me3	Kme3STGGKAPR	0.10 $\pm$ 0.02	0.17 $\pm$ 0.03	0.22 $\pm$ 0.02
H3.K18_K23_noPTM	KQLATKAAR	60.62 $\pm$ 0.93 *	62.46 $\pm$ 0.11	62.72 $\pm$ 0.83
H3.K18ac	KacQLATKAAR	24.33 $\pm$ 0.30 **	22.82 $\pm$ 0.25	22.17 $\pm$ 0.57
H3.K23ac	KQLATKacAAR	6.90 $\pm$ 0.35	6.82 $\pm$ 0.08	7.14 $\pm$ 0.22
H3.K18ac_K23ac	KacQLATKacAAR	8.15 $\pm$ 0.40	7.90 $\pm$ 0.09	7.97 $\pm$ 0.25
H3.1.K27_K36_K37_noPTM	KSAPATGGVKKPHR	10.19 $\pm$ 1.33	9.55 $\pm$ 1.09	9.22 $\pm$ 0.83
H3.1.K27ac	KacSAPATGGVKKPHR	0.18 $\pm$ 0.01	0.17 $\pm$ 0.02	0.17 $\pm$ 0.01
H3.1.K36ac	KSAPATGGVKacKPHR	0.15 $\pm$ 0.01	0.14 $\pm$ 0.01	0.13 $\pm$ 0.01
H3.1.K27ac_K36me2	KacSAPATGGVKme2KPHR	0.08 $\pm$ 0.01	0.07 $\pm$ 0.01	0.09 $\pm$ 0.03
H3.1.K27ac_K36me3	KacSAPATGGVKme3KPHR	0.78 $\pm$ 0.14	0.76 $\pm$ 0.09	0.63 $\pm$ 0.03
H3.1.K27me2_K36ac	Kme2SAPATGGVKacKPHR	0.16 $\pm$ 0.02	0.16 $\pm$ 0.02	0.14 $\pm$ 0.02
H3.1.K27me3_K36ac	Kme3SAPATGGVKacKPHR	0.08 $\pm$ 0.04	0.07 $\pm$ 0.01	0.08 $\pm$ 0.04
H3.1.K27me1	Kme1SAPATGGVKKPHR	49.45 $\pm$ 3.87	49.40 $\pm$ 3.38	43.24 $\pm$ 0.26
H3.1.K27me2	Kme2SAPATGGVKKPHR	18.68 $\pm$ 1.87	19.48 $\pm$ 1.46	24.08 $\pm$ 0.77 *
H3.1.K27me3	Kme3SAPATGGVKKPHR	6.53 $\pm$ 1.63	6.68 $\pm$ 1.28	7.74 $\pm$ 0.24
H3.1.K36me1	KSAPATGGVKme1KPHR	2.27 $\pm$ 0.36	2.32 $\pm$ 0.29	2.85 $\pm$ 0.78
H3.1.K36me2	KSAPATGGVKme2KPHR	1.49 $\pm$ 0.05	1.41 $\pm$ 0.08	1.63 $\pm$ 0.26
H3.1.K36me3	KSAPATGGVKme3KPHR	9.97 $\pm$ 1.76	9.78 $\pm$ 1.43	10.22 $\pm$ 0.59

Abundance of histone methylation and acetylation marks on histone H3 in 4-week-old *Arabidopsis* plants as determined by LC-MS. Relative abundance of 31 PTMs involving lysine acetylation and methylation marks on histone H3 in 4-week-old rosette leaves grown under long-day conditions and harvested 5 h after the day-time start from wt, *sahh1*, and *gsnor1-3* plants. Statistics: values are the relative abundance of each histone motif at each peptide and represent the mean $\pm$ SD ($n = 3$). * ($p < 0.05$), ** ($p < 0.01$), and *** ($p < 0.001$) represent significant differences between wt and mutant lines (ANOVA, Dunnett's multiple comparisons test). Statistical analysis was performed with GraphPad Prism version 7.05. For calculation of motif abundance, refer to Feller et al. [76]. The motif identifier name contains the PTM type and position. H3.K4me1: abundance of mono-methylation on K4 relative to H3.K4me2, H3.K4me3, and H3.K4noPTM. Kac, lysine acetylation; Kme1, lysine mono-methylation; Kme2, lysine di-methylation; Kme3, lysine tri-methylation; noPTM, peptide without PTM.

The altered H3K9me2 levels were confirmed by immunoblotting using an anti-H3K9me2 antibody (Figure 2 A,B). In addition, the H3.1K27me2 mark was significantly increased by 23% in *gsnor1-3* plants.

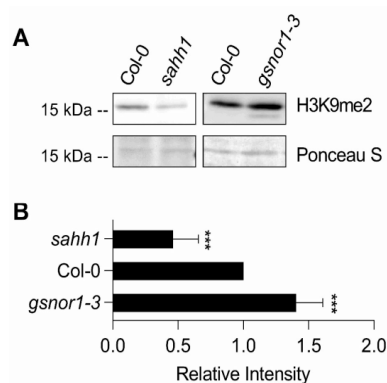


Figure 2. Histone H3K9me2 methylation level is altered in *gsnor1-3* and *sahh1*. (A) H3K9me2 immunoblot. Histones were acid extracted from 4-week-old rosette leaves grown under long-day conditions and harvested 5 h after the day-time start and probed against H3K9me2 marks by immunoblotting. As the loading control, the Ponceau S-stained membrane is shown. One representative experiment is shown. (B) Quantification of immunoblot results. Signal intensities were measured using ImageJ software and normalized to the amount of loaded H3. Statistics: values are expressed as fold change over wt and represent the mean $\pm$ SD of at least three independent experiments ($n = 4-7$). Grubb's outlier test ($\alpha = 0.05$) was performed. *** ($p < 0.001$) represents significant differences between wt and mutant lines (ANOVA, Dunnett's multiple comparisons test). Statistical analysis was performed with GraphPad Prism version 7.05.

3.3. SAHH1 and GSNOR1 Functions Affect DNA Methylation

Since H3K9me2 is functionally linked to DNA methylation [43,83,84], we postulated that the observed altered global H3K9me2 level in *sahh1* and *gsnor1-3* plants would entail changes in DNA methylation.

We used the *A. thaliana* Col-0 *TS-GUS* (L5, 6b5) line, which possesses a transcriptionally silent highly repetitive GUS transgene on chromosome III [64], to analyze the effect of GSNOR1 and SAHH1 on DNA methylation. Transcriptional gene silencing (TGS) is generally concomitant with high levels of DNA methylation and inactive chromatin marks such as H3K9me2. We crossed the *TS-GUS* (L5) line with *sahh1* and *gsnor1-3* mutant lines (Supplemental Figure S3) and assessed the reactivation of *TS-GUS* in 10-day-old seedlings (Figure 3). As a control, seedlings were grown in the presence of DHPA, an SAHH-specific inhibitor previously demonstrated to reactivate *TS-GUS* [53]. DHPA induced the reactivation of the GUS transgene in each mutant background. However, in the absence of DHPA, activation of the GUS transgene was only observed in the *sahh1* background, but not in the *gsnor1-3* background. These results demonstrate a hypomethylation phenotype of *sahh1* and argue for an unchanged methylation status or a hypermethylation phenotype of *gsnor1-3*.

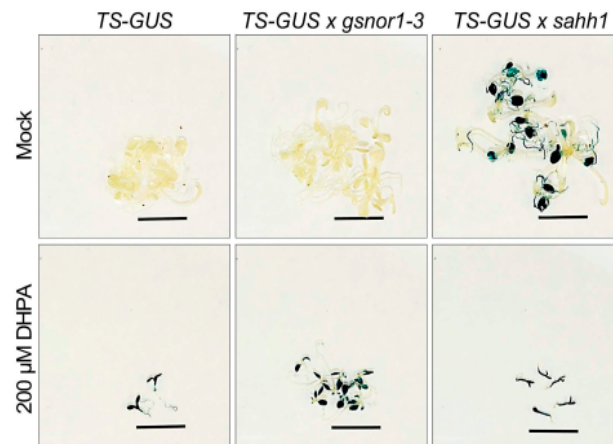


Figure 3. Reactivation of *TS-GUS* in *gsnor1-3* and *sahh1*. Blue staining indicates release of gene silencing. Plantlets were grown in liquid 1× MS under short-day conditions supplemented with water (mock) as control, or seedlings were grown in the presence of 200 μM DHPA. Due to DHPA, reduced growth is observed as previously demonstrated [53]. GUS reactivation was visible in *sahh1* but not in the *gsnor1-3* background. Treatments with the SAHH inhibitor DHPA released *TS-GUS* silencing in all mutant backgrounds. Scale bar = 1 cm.

DNA methylation in *sahh1* and *gsnor1-3* was further analyzed by whole-genome bisulfite sequencing (WGBS). WGBS allows studying genome-wide DNA methylation at single-nucleotide resolution. To assess the bisulfite conversion efficiency, reads were mapped to the non-methylated chloroplast genome, resulting in an average conversion rate of more than 98%. The mean methylation levels per DNA methylation context (CG, CHG, or CHH, where H = A, C, or T) are listed in Table 2.

Table 2. Mean methylation rates per context (±SD) as analyzed by WGBS ($n = 2$). Mean methylation rates in CG, CHG, or CHH context (H = A, C, or T) in Col-0/wt, *sahh1*, and *gsnor1-3* were calculated from cytosines that were covered by at least 5 reads.

	CG	CHG	CHH
Col-0/wt	22.90 ± 2.45%	6.38 ± 1.28%	1.54 ± 0.27%
<i>gsnor1-3</i>	20.32 ± 0.19%	5.59 ± 0.07%	1.38 ± 0.03%
<i>sahh1</i>	19.43 ± 1.18%	3.11 ± 0.26%	0.95 ± 0.10%

These data are in accordance with the average methylation levels of 24% CG, 7% CHG, and 2% CHH found in *Arabidopsis* [85]. The mean methylation rates in *gsnor1-3* are similar to those in wt, whereas *sahh1* shows decreased methylation rates (Table 2). However, at the level of chromosomal distribution, hypermethylation in *gsnor1-3* was most pronounced in the TE-rich pericentromeric regions in the CHG context (Figure 4). For *sahh1*, we observed the strongest effect in the CHG context, followed by CHH and CG compared to wt methylation rates (Table 2). Loss of methylation in *sahh1* was unevenly distributed along the chromosomes and was most pronounced in the highly methylated TE-rich pericentromeric regions, particularly for CHG and CHH (Figure 4). Taken together, DNA methylation is altered in both mutants compared to wt.

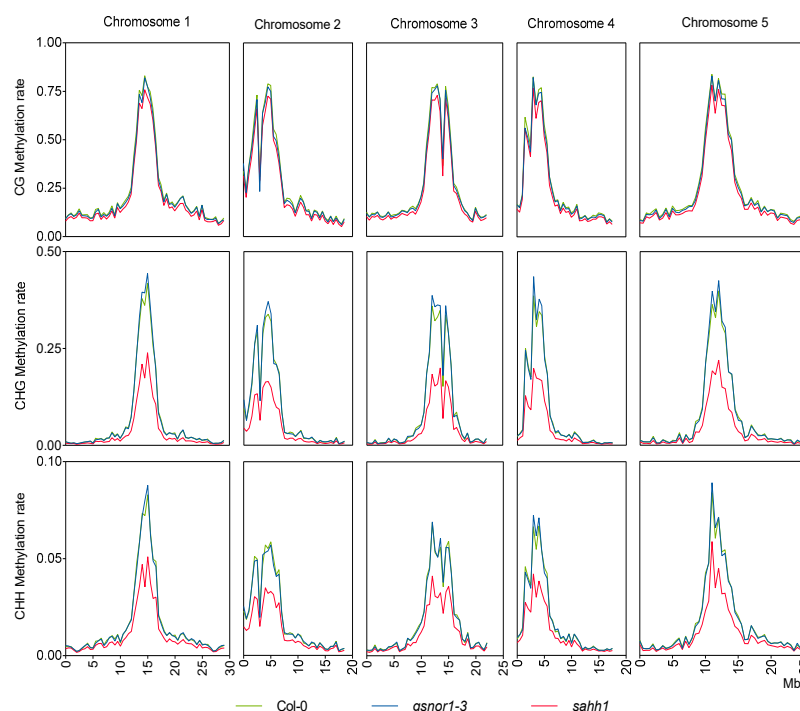


Figure 4. Chromosomal distribution of DNA methylation is altered in *gsnor1-3* and *sahh1*. The methylation levels across the chromosomes in each sequence context were calculated with MethGeno [86] for each replicate. Then, replicates were merged, and graphs were made with GraphPad Prism. Average methylation of all cytosines within a 0.5 Mbp interval is plotted.

3.4. GSNOR1 and SAHH1 Regulate DNA Methylation of TEs and Genes

To assess whether GSNOR1 and SAHH1 affect the methylation status of the defined genomic regions, we first called methylation regions (MRs) using the adaptation of a two-state hidden Markov model-based approach and identified differentially methylated regions (DMRs) in pairwise comparisons (*gsnor1-3* vs. wt, and *sahh1* vs. wt) according to Hagmann et al. [70]. We identified 42,304 and 40,305 MRs in wt and *gsnor1-3*, respectively. Comparing wt and *sahh1* resulted in 42,288 and 51,223 identified MRs, respectively. DMR identification in pairwise comparisons (mutant vs. wt) revealed 752 and 292 DMRs for *sahh1* and *gsnor1-3* (Supplemental Table S2), respectively. In *sahh1*, 35 DMRs were hypermethylated and 717 were hypomethylated relative to wt (Figure 5A), in line with the overall decrease in *sahh1* (Table 2). In contrast, *gsnor1-3* showed considerably more hypermethylated (231) than hypomethylated (61) DMRs relative to wt (Figure 5B), despite the fact that mean methylation rates were similar to those of wt (Table 2). In summary, *sahh1* and *gsnor1-3* mutants predominantly showed local hypo- and hypermethylation, respectively.

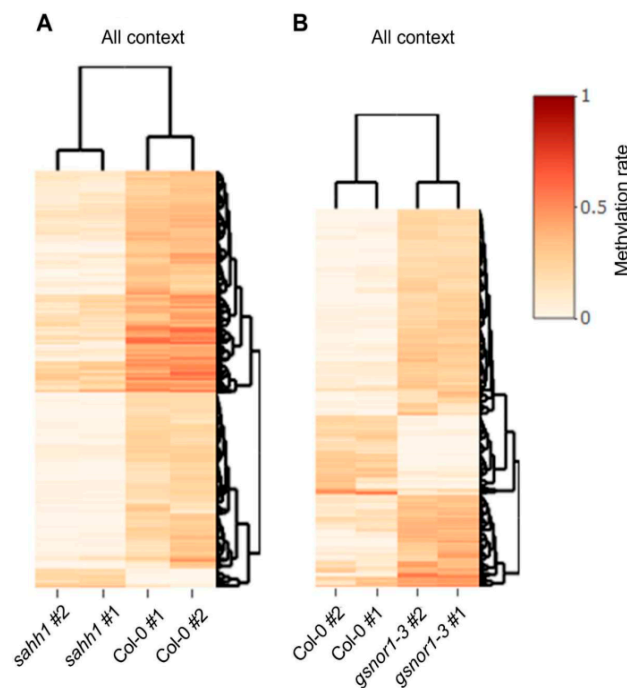


Figure 5. Enrichment of hypo- and hyper-DMRs in *sahh1* and *gsnor1-3*, respectively. Heatmaps of hierarchically clustered DMRs identified in pairwise comparisons of wt vs. *sahh1* methylome (A) and wt vs. *gsnor1-3* methylome (B). DNA was extracted from 4-week-old rosette leaves grown under long-day conditions, harvested 5 h after the day-time start, and subjected to WGBS. Heatmaps represent the methylation level across DMRs: red = 100% methylated, white = 0% methylated. Two biological replicates were analyzed for each genotype.

Genomic feature annotation showed that DMRs mainly mapped to the genic (CDS), 3kb up- or downstream flanking regions of genes (hereafter, genes overlapping with identified DMRs in their genic, 3kb up- and/or downstream region are summarized as differentially methylated genes (DMGs)) and TEs (Supplementary Figure S4A,B).

In detail, loss of the GSNOR1 function resulted in an enrichment of hypermethylated DMGs and TEs (Figure 6A). In total, 587 DMGs were identified in *gsnor1-3* (Supplemental Table S3). Among those, 449 were hypermethylated and 138 were hypomethylated. DMGs with DMRs in multiple genomic elements were identified, as illustrated in the Venn diagram (Figure 6B). For instance, hypermethylation of AT5G46295 encoding a transmembrane protein is observed in its 3kb upstream flanking and genic region (Figure 6C). Loss of the GSNOR1 function resulted in an enrichment of hypermethylated TEs (Figure 6A,D). In detail, 55 and 12 TEs overlapping with hyper- and hypomethylated DMRs were identified (Supplemental Table S4), respectively. TEs classified as retrotransposons in the superfamilies LTR/Gypsy and LINE/L1, as well TEs classified as DNA transposons belonging to the superfamilies MuDR and RC/Helitron, were mainly hypermethylated in *gsnor1-3* (Figure 6D). A snapshot in the EPIC-CoGE browser of a representative TE (AT3TE65465, LTR/Gypsy) overlapping with a hyper-DMR is shown in Figure 6E.

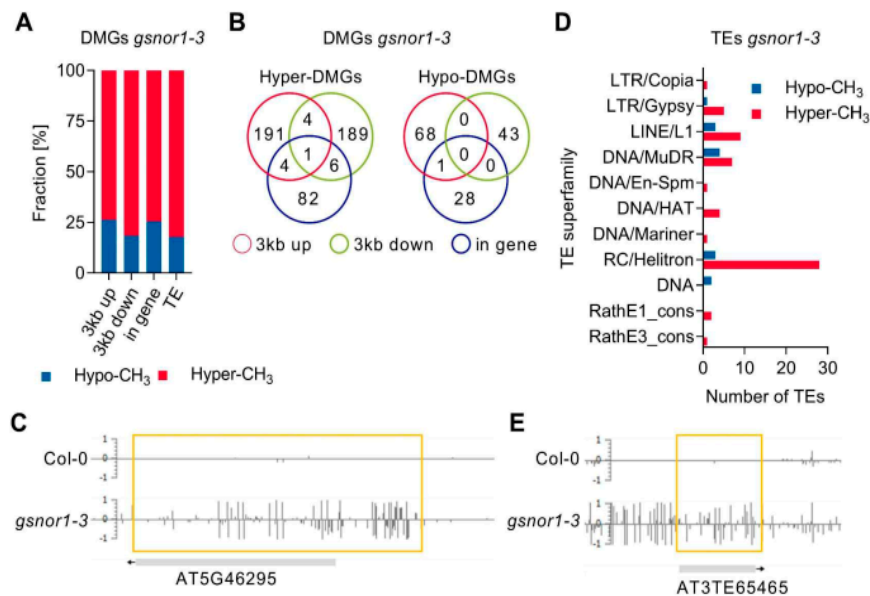


Figure 6. Loss of GSNOR1 function results in an enrichment of hypermethylated DMGs and TEs. (A) Stacked bar plot showing the fraction of hyper- and hypo-DMGs and TEs with respect to DMRs found in their genic, 3kb up- or downstream region or TE coding region. (B) Venn diagram summarizing DMGs with DMRs in multiple genomic features (genic, 3kb up- and/or downstream flanking regions). (C) Snapshot of AT5G46295 in the EPIC-CoGE browser. (D) Distribution of differentially methylated TEs over TE superfamilies. (E) Snapshot of AT3TE65465 in the EPIC-CoGE browser. DNA methylation data have been uploaded to the epigenome browser of EPIC (EPIC-CoGE) by Prof. Dr. Claude Becker (ID 2234 unpublished). DNA methylation analysis was performed in duplicates, and average methylation ratios calculated in the CoGE browser are shown. TE classification according to www.arabidopsis.org accessed on 24 December 2019.

SAHH1 knock-down resulted mainly in an enrichment of hypomethylated DMGs and TEs (Figure 7A). In *sahh1* plants, 1299 DMGs were identified (Supplemental Table S5). Among those, 72 are hypermethylated and 1227 are hypomethylated. Of note, three of those DMGs in *sahh1* possess hyper- and hypo-DMRs (At1g65220, At3g54730, At4g13440). As illustrate in the Venn diagrams, DMGs with DMRs in multiple genomic elements were identified (Figure 7B). For instance, AT3G50250 encoding a transmembrane protein is hypomethylated in its 3kb upstream flanking and genic region (Figure 7C).

TEs were mainly hypomethylated in *sahh1* (Figure 7A,D; Supplemental Table S6). In detail, 3 TEs and 271 TEs with hyper- and hypomethylated DMRs were identified, respectively. Hypomethylation was mainly found in members of the retrotransposon superfamilies LTR/Copia and LINE/L1, and in members of the DNA transposon superfamilies MuDR and RC/Helitron (Figure 7D). A snapshot in the EPIC-CoGE browser of a representative hypomethylated TE (AT1TE93270, DNA/HAT) is shown in Figure 7E. Gene Ontology enrichment analysis of DMGs identified in *gsnor1-3* and *sahh1* did not result in significantly enriched GO terms.

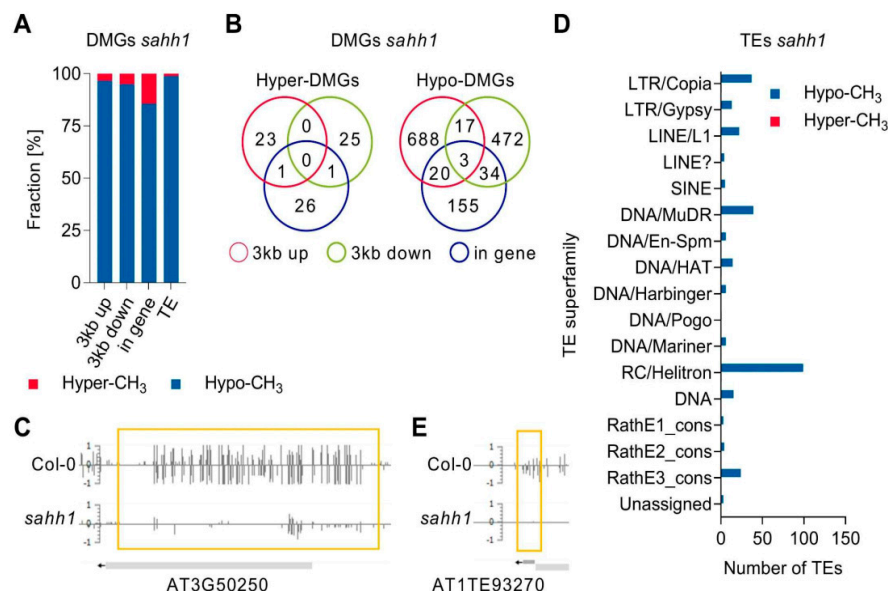


Figure 7. Knock-down of SAHH1 function results in an enrichment of hypomethylated DMGs and TEs. (A) Stacked bar plot showing the fraction of hyper- and hypo-DMGs and TEs with respect to DMRs found in their genic, 3kb up- or downstream region or TE coding region. (B) Venn diagram summarizing DMGs with DMRs in multiple genomic features (genic, 3kb up- and/or downstream flanking regions). (C) Snapshot of AT3G50250 in the EPIC-CoGE browser. (D) Distribution of differentially methylated TEs over TE superfamilies. (E) Snapshot of AT1TE93270 in the EPIC-CoGE browser. DNA methylation analyses were performed in duplicates, and average methylation ratios calculated in the CoGE browser are shown. TE classification according to www.arabidopsis.org (24 December 2018).

3.5. Transcriptomic Profiling of *gsnor1-3* and *sahh1* Plants

We performed RNA sequencing to link the observed differential DNA methylation in *sahh1* and *gsnor1-3* with gene expression and physiological functions. Among the 1129 differentially expressed genes (DEGs; $|\log_2FC| > 1$, adjusted p -value less than 0.1) identified in *gsnor1-3* relative to wt, three quarters (949) were downregulated, and only 180 were upregulated (Figure 8A,B; Supplemental Table S7); similarly, most TE families were downregulated (Figure 8C; Supplemental Table S8). Transcriptional profiling of *sahh1* revealed 394 DEGs which were evenly up- and downregulated (211 vs. 183) (Figure 8D,E; Supplemental Table S9). In contrast, expression of TEs from most TE families was upregulated in *sahh1* (Figure 8D,F; Supplemental Table S10).

RNA was extracted from 4-week-old rosette leaves grown under long-day conditions and harvested 5 h after the day-time start ($n = 4$). Significant criteria: $|\log_2FC| > 1$, adjusted p -value less than 0.1. Gene Ontology (GO) term enrichment analysis revealed that among the most significantly enriched molecular functional categories of the upregulated genes in *gsnor1-3* were the “catalytic activity”, “glutathione transferase”, “glycosyltransferase”, and “oxidoreductase activity” categories (Supplemental Table S11). Moreover, biological process categories such as “response to light”, “response to UV”, “cellular response to reactive oxygen species”, “cellular response to oxidative stress”, and “cellular response to nitric oxide” were significantly enriched. According to that, the transcript profile analysis of *gsnor1-3* plants suggests a pre-induced antioxidant system under normal growth conditions, as previously reported for *gsnor* (Ws background [39]). Further, •NO response was enriched among upregulated genes. This is in agreement with the enhanced RSNO/•NO level in

gsnor1-3 [34,35]. The majority of downregulated genes in the GO term “biological process” are related to the defense response and response to stress.

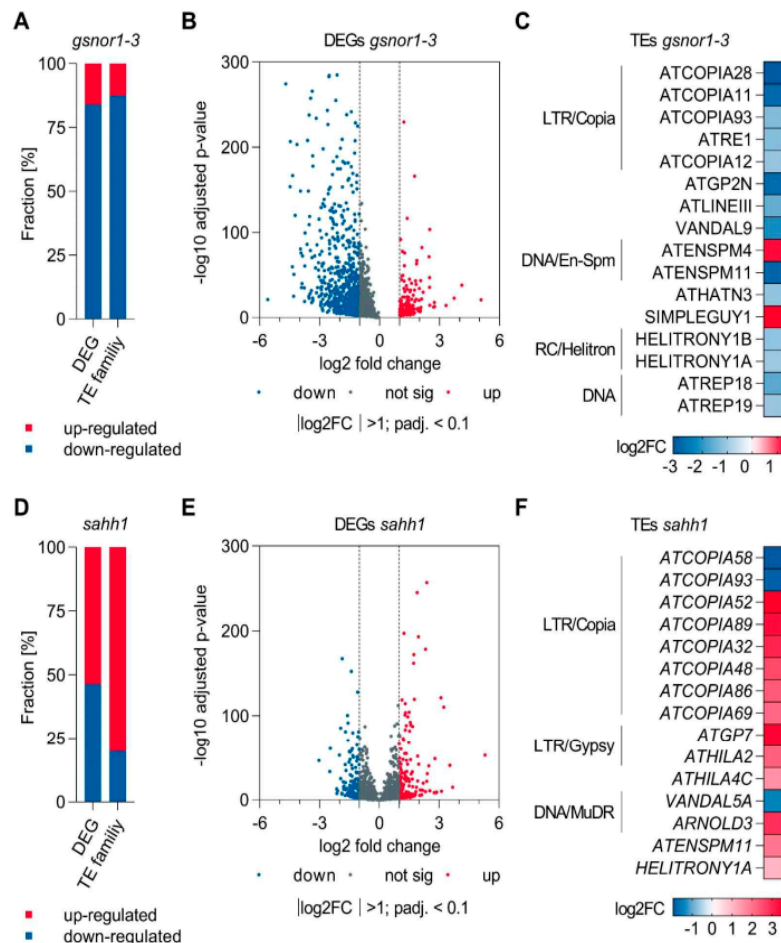


Figure 8. Mutations in *GSNOR1* and *sahh1* result in transcriptional reprogramming. (A,D) Stacked bar plots showing the fraction of significant differentially expressed genes (DEGs) and transposable element (TE) families in *gsnor1-3* and *sahh1*. (B,E) Volcano plot highlighting significant DEGs in *gsnor1-3* and *sahh1*. Blue and red dots mark significantly decreased or increased expressed genes, respectively. (C,F) Heat map showing differential expression of TE families in *gsnor1-3* and *sahh1*. If more than one family within a superfamily is differentially expressed, superfamilies are indicated. TEs are classified according to www.arabidopsis.org (accessed on 24 December 2018).

Concerning *sahh1*, the terms “DNA-binding transcription factor activity” and “metal ion binding” were the dominant categories among the molecular functions enriched for downregulated genes (Supplemental Table S12). Among biological processes, terms related to “hormones” and “response to chemical” were over-represented. For instance, *LIPOXYGENASE 4* involved in the biosynthesis of the plant hormone jasmonic acid was downregulated, as previously reported [56]. Further, the term anthocyanin-containing

compound biosynthesis was found when analyzing upregulated DEGs in *salh1*, which is in line with previous studies [56].

3.6. Integrative Analysis of WGBS and RNA-Seq Data

To test whether differential methylation was associated with differential expression, we looked for associations between the WGBS and RNA-seq datasets (DEG–DMG candidates). The integrative analysis of DMGs and DEGs at the gene level revealed that about 4% of DMGs were differentially expressed (percentages are relative to DMGs). Hypo- and hypermethylation were positively and negatively correlated with transcription (Table 3). Loss of GSNOR function resulted in hypermethylation of up- and downregulated TEs in the CHG context (Supplemental Figure S5). Integrative analysis of differentially methylated TEs and differentially expressed TE families revealed that DNA methylation is negatively associated with TE expression. In detail, TE families ATLINEIII, ATHATN3, and HELITRON1A were downregulated (expression analysis performed at family level; Supplemental Table S8), and members of those TE families were hypermethylated (Supplemental Table S4).

Table 3. Integrative analysis of DMGs and DEGs in *gsnor1-3*.

Overlap of Significantly Downregulated Genes with DMGs								
Chr	Start	bp	CH ₃	Feature	Gene ID	log2FC	padj.	Description
Chr5	8751681	42	+	3kb down	AT5G25250	−3.37	2.0×10^{-266}	Flotillin-like protein 1 (UV-stress)
Chr5	9309455	206	−	3kb up	AT5G26690	−3.35	2.3×10^{-23}	Heavy metal-associated isoprenylated plant protein 2 (stress response; but not much data available)
Chr2	11812888	185	+	3kb down	AT2G27690	−2.94	7.6×10^{-36}	Cytochrome P450 94C1 (jasmonoyl-L-isoleucine; wounding)
Chr5	8751681	42	+	3kb up	AT5G25260	−2.80	2.0×10^{-64}	Flotillin-like protein 2 (UV-stress)
Chr2	9741371	43	+	in gene	AT2G22880	−2.24	3.8×10^{-12}	At2g22880 (Hypoxia, UV-stress, salt, wounding)
Chr2	15110344	63	+	3kb up	AT2G35980	−2.07	1.7×10^{-8}	NDR1/HIN1-like protein 10 (Hypoxia, salt, biotic stress)
Chr2	18325130	77	+	3kb up	AT2G44380	−2.05	1.8×10^{-8}	At2g44380 (biotic stress)
Chr1	24395763	100	+	3kb up	AT1G65610	−1.82	4.4×10^{-6}	Endoglucanase 7 (biotic stress)
Chr5	5767502	32	−	3kb up	AT5G17490	−1.56	6.8×10^{-10}	DELLA protein RGL3 (wounding, cold)
Chr3	22556563	37	+	3kb down	AT3G60966	−1.45	1.8×10^{-4}	RING/U-box superfamily protein
Chr2	18325130	77	+	3kb up	AT2G44400	−1.39	1.1×10^{-3}	Cysteine/Histidine-rich C1 domain family protein
Chr1	27068879	85	+	3kb down	AT1G71910	−1.34	7.2×10^{-6}	At1g71910
Chr2	3304271	210	−	in gene	AT2G07774	−1.31	4.1×10^{-5}	unknown protein
Chr5	18136940	44	+	3kb up	AT5G44920	−1.21	2.5×10^{-3}	TIR domain-containing protein
Chr5	18136984	64	+	3kb up	AT5G44920	−1.21	2.5×10^{-3}	TIR domain-containing protein
Chr3	3063382	181	−	in gene	AT3G09960	−1.21	4.7×10^{-3}	Calcineurin-like metallo-phosphoesterase superfamily protein
Chr5	18779966	240	+	in gene	AT5G46295	−1.19	2.3×10^{-4}	Transmembrane protein
Chr5	18780206	180	+	3kb up	AT5G46295	−1.19	2.3×10^{-4}	Transmembrane protein
Chr2	12426536	39	−	3kb down	AT2G28940	−1.17	1.7×10^{-17}	At2g28940
Chr1	4123656	44	−	3kb up	AT1G12160	−1.13	2.0×10^{-4}	Flavin-containing monooxygenase FMO GS-OX-like 1
Chr1	21823145	288	−	3kb down	AT1G59124	−1.08	2.9×10^{-19}	Probable disease resistance protein RF45

Table 3. Cont.

Overlap of Significantly Upregulated Genes with DMGs								
Chr	Start	bp	CH ₃	Feature	Gene Id	log2FC	padj.	Description
Chr5	7376314	54	–	3kb up	AT5G22300	1.82	1.3×10^{-39}	Bifunctional nitrilase/nitrile hydratase NIT4 (UV stress, biotic stress)
Chr3	9173846	95	–	3kb up	AT3G25190	1.30	2.7×10^{-4}	Vacuolar iron transporter homolog 2.1 (biotic stress)
Chr5	17145940	99	+	3kb up	AT5G42760	1.29	1.8×10^{-8}	Leucine carboxyl methyltransferase
Chr4	13766210	58	+	3kb up	AT4G27570	1.02	1.1×10^{-2}	UDP-glycosyltransferase 79B3 (jasmonate, cold)
Chr5	9637396	186	+	3kb up	AT5G27330	1.02	2.3×10^{-11}	Prefoldin chaperone subunit family protein

The methylation status in *gsnor1-3* compared to wt is given as (–) and (+) referring to hypo- and hypermethylation, respectively. DMRs are annotated with genomic features (3kb up- or downstream, and in gene). Abbreviations: Chr, chromosome; start, DMR start position; bp, length of overlapping DMR with genomic feature. Statistics for RNA-seq: $|\log_2FC| > 1$, adjusted *p*-value less than 0.1.

Metaplot analysis revealed that the DNA methylation levels of the identified up- and downregulated genes in *sahh1* tended to decrease compared to wt (Supplemental Figure S6). An integrative analysis of DMGs and DEGs at the gene level revealed that about 1.7% of DMGs were differentially expressed (percentages are relative to DMGs). Upregulated genes mainly correlated with decreased methylation in their 3kb upstream and genic region (Supplemental Figure S6). Interestingly, downregulated genes were also associated with reduced DNA methylation levels in their 3kb up- or downstream region in *sahh1* (Table 4; Supplemental Figure S6). Hypomethylation in each sequence context was observed in up- and downregulated TEs in *sahh1* (Supplemental Figure S6).

Integrative analysis of differentially methylated TEs and differentially expressed TE families revealed that DNA hypomethylation caused TE activation. In detail, TE families ATCOPIA89, ATHILA2, and HELITRON1A were upregulated (expression analysis performed at family level; Supplemental Table S10), and members of those TE families possess hypo-DMRs (Supplemental Table S6). Taken together, there is a low correlation between altered DNA methylation and the expression of protein-coding genes, whereas DNA methylation is principally negatively correlated with TE expression.

Table 4. Integrative analysis of DMGs and DEGs in *sahh1*.

Overlap of Significantly Downregulated Genes with DMGs								
Chr	Start	bp	CH ₃	Feature	Gene ID	log2FC	padj.	Description
Chr5	19178939	108	–	3kb up	AT5G47230	–1.94	1.3×10^{-25}	Ethylene responsive element binding factor 5 ERF5
Chr5	5907343	107	–	3kb up	AT5G17860	–1.88	1.3×10^{-8}	Cation/calcium exchanger 1
Chr2	18497356	377	–	3kb down	AT2G44840	–1.79	9.7×10^{-16}	Ethylene-responsive transcription factor 13 ERF13
Chr1	13837861	133	–	3kb up	AT1G36622	–1.51	4.6×10^{-5}	Transmembrane protein
Chr1	13837994	23	–	3kb up	AT1G36622	–1.51	4.6×10^{-5}	Transmembrane protein
Chr5	7261113	306	–	3kb up	AT5G21960	–1.26	2.2×10^{-3}	Ethylene-responsive transcription factor ERF016
Chr5	16023667	82	–	3kb up	AT5G40010	–1.24	8.2×10^{-4}	AAA-ATPase ASD, mt
Chr1	26140005	248	–	3kb up	AT1G69530	–1.08	1.7×10^{-128}	Expansin
Chr4	14031509	89	–	3kb down	AT4G28350	–1.07	8.3×10^{-10}	Probable L-type lectin-domain containing receptor kinase VII.2

Table 4. Cont.

Chr	start	bp	CH ₃	Feature	Gene ID	log2FC	padj.	Description
Chr5	9206475	54	–	in gene	AT5G26270	5.31	3.0×10^{-53}	unknown protein
Chr3	20260251	114	–	in gene	AT3G54730	3.69	1.3×10^{-15}	Putative transmembrane protein At3g54730
Chr3	20260365	94	–	in gene	AT3G54730	3.69	1.3×10^{-15}	Putative transmembrane protein At3g54730
Chr3	20260459	7	+	3kb down	AT3G54730	3.69	1.3×10^{-15}	Putative transmembrane protein At3g54730
Chr5	18208166	230	–	in gene	AT5G45095	3.12	7.8×10^{-11}	Putative uncharacterized protein
Chr1	12851246	141	–	3kb up	AT1G35140	2.38	1.2×10^{-257}	Protein EXORDIUM-like 1
Chr4	6431517	56	+	in gene	AT4G10380	1.44	6.9×10^{-10}	At4g10380
Chr2	12887310	93	–	3kb down	AT2G30210	1.44	9.8×10^{-4}	Laccase-3
Chr2	13160854	47	–	3kb up	AT2G30930	1.43	2.3×10^{-99}	Expressed protein
Chr1	3980123	55	–	in gene	AT1G11785	1.31	2.6×10^{-3}	Putative uncharacterized protein
Chr3	9173846	95	–	3kb up	AT3G25190	1.29	6.6×10^{-4}	Vacuolar iron transporter homolog 2.1
Chr3	21509510	77	–	3kb up	AT3G58070	1.14	2.0×10^{-3}	GIS
Chr3	20206910	10	–	3kb up	AT3G54580	1.01	4.4×10^{-2}	Proline-rich extensin-like family protein

The methylation status in *inasah1* compared to wt is given as (–) and (+) referring to hypo- and hypermethylation, respectively. DMRs are annotated with genomic features (3kb up- or downstream, and in gene). Abbreviations: Chr, chromosome; start, DMR start position; bp, length of overlapping DMR with genomic feature. Statistics for RNA-seq: $|\log_2FC| > 1$, adjusted *p*-value less than 0.1.

4. Discussion

4.1. Loss of GSNOR1 Function Results in an Increased Methylation Index

The main function of the methylation cycle is to produce SAM for transmethylation reactions and to recycle the by-product inhibitor SAH [48,49]. The SAM/SAH ratio (MI) is considered as a metabolic indicator of the organismal methylation status [51], since SAM is used as a methyl donor by methyltransferases, and SAH competitively inhibits most of the known SAM-dependent methyltransferases [87]. Loss of GSNOR1 triggered a metabolic reprogramming affecting the methylation cycle by increasing the level of SAM. Since the level of SAH is not altered in *gsnor1-3*, the SAM/SAH ratio consequently increased (Figure 1A,C). In *sahh1*, the level of SAM is also enhanced, but since SAH accumulates stronger, the SAM/SAH ratio is finally decreased in *sahh1*. Surprisingly, metabolites of pathways connected to the methylation cycle (MTA, Cys, GSH) are increased in both plant mutants, concluding that the GSNOR and SAHH1 function is involved in regulating the levels of these metabolites, which also influence methylation processes. In terms of epigenetics, GSH was demonstrated to impact epigenetic mechanisms in the animal system [88]. For instance, the activity of the liver isoform SAMS1 depends on the GSH/GSSG ratio [88], indicating a crosstalk between GSH/GSSG levels and SAM synthesis. Moreover, SAM inhibits demethylase activity in vitro and in cells [89]. However, since SAM is highly unstable, it is not clear whether its in vivo activities are caused by SAM or by SAM metabolites, such as MTA [90]. MTA was shown to affect histone methylation as a histone methyltransferase inhibitor [91]. Furthermore, the combination of metabolic changes might have synergistic effects on the epigenetic landscape.

Interestingly, transcriptomic changes of genes involved in the methylation cycle were not observed in *gsnor1-3* (Supplemental Table S7). We confirmed in vitro S-nitrosation of SAHH1 by GSNO using purified recombinant SAHH1 and plant protein extracts (Supplemental Figure S2A,C). Furthermore, other groups demonstrated that S-nitrosation strongly inhibits SAHH1 activity in vitro [82]. This, at least, raises the possibility that the formation of SAHH1-SNO plays a role in fine-tuning the SAHH1 enzyme activity in respect to epigenetic methylation marks under yet unknown conditions. However, the S-nitrosation of SAHH1 and its influence on the enzyme activity in vivo would certainly require further experimental analysis.

Interestingly, metabolites of pathways connected to the methylation cycle, such as MTA, Cys, and GSH, were increased in both *gsnor1-3* and *sahh1* (Figure 1D–F), demonstrating that GSNOR and SAHH1 are also important for regulating the levels of these metabolites.

4.2. GSNOR1 Function Is Crucial for the Maintenance of Histone Methylation and DNA Methylation

Several lines of evidence have demonstrated that an altered MI affects histone and DNA methylation in plants and animals ([42,50,51] and references therein). To date, the interconnection between an increased MI and hypermethylation has been rarely reported [92,93], whereas a decreased MI concomitant with a hypomethylated phenotype, as observed in the *sahh1* plants, has been described frequently ([50,51] and references therein). Indeed, a decreased MI predominantly results in loss of H3K9me2 and loss of non-CG methylation, whereas other histone methylation marks, such as H3K27me1 and H3K9me1, and CG methylation are less affected in *Arabidopsis* ([50,51] and references therein).

Loss of the GSNOR1 function results in global hypermethylation of H3K9me2 and H3.1.K27me2 (Table 1). However, we can only speculate about the exact GSNOR/•NO-dependent molecular mechanisms regulating the methylation of these histone marks. Besides modulation of the methylation cycle via affecting SAM levels (Figure 1A), GSNOR/•NO could also regulate the expression of genes encoding for histone modifiers and/or could directly modulate the activity of proteins involved in histone (de-)methylation via GSNOR/•NO-mediated PTMs. Indeed, several genes involved in histone methylation are differentially expressed in *Arabidopsis* plants with impaired •NO homeostasis or after •NO donor treatment (summarized in Lindermayr et al. [94]). Moreover, regulation of the *Arabidopsis* histone arginine demethylase PRMT5 by S-nitrosation has been reported [95].

Histone methylation is functionally linked to DNA methylation. For instance, H3K9me2 and non-CG methylation are connected by a reinforcing loop, which perpetuates both epigenetic marks catalyzed by CMT2/3 and SUVH4/5/6, respectively [84]. However, SUVH5 and SUVH6 can also bind to DNA that is methylated in the CG context in vitro [96], supporting the view that CG methylation also contributes to H3K9me2 deposition. In addition to CG methylation, it has been known for many years that CHH methylation generated by the RNA-directed DNA methylation pathway is also involved in H3K9me2 deposition [97–100]. Profiling of cytosine methylation patterns with high resolution by WGBS demonstrated that loss of the GSNOR1 function affects DNA methylation. Although the MI was increased, almost unchanged global mean DNA methylation rates were observed in *gsnor1-3* (Table 2). However, in relation to the genome-wide position of methylated cytosines, rather hyper-DMRs, as opposed to hypo-DMRs, were identified (Figure 5B). In fact, the number of hyper-DMRs was more than 3.8 times that of hypo-DMRs (61 hypo-DMRs; 231 hyper-DMRs). This finding indicates that the GSNOR1 function seems to be important for the hypermethylation of these regions. •NO-induced changes in the expression of genes related to DNA methylation further demonstrated the importance of •NO for DNA (de-)methylation processes (summarized in Lindermayr et al. [94]; Supplemental Table S7). Based on studies in the human/animal field, different effects could contribute to the altered DNA methylation pattern in *gsnor1-3*. For instance, reduced active DNA demethylation could tilt the equilibrium of methylation processes toward methylation in *gsnor1-3*. In this context, elevated levels of SAM, as observed in *gsnor1-3* (Figure 1A), counteract active DNA demethylation in human cells [89,101]. Further, mammalian TET enzymes involved in DNA demethylation are inhibited by •NO due to the formation of a nitrosyl–iron complex with their catalytic iron [102]. Similarly, the iron–sulfur-containing ROS1/DME DNA demethylases [103] could be affected by •NO in *gsnor1-3*. The attack of iron–sulfur clusters by •NO [104] has been previously shown. For instance, •NO inhibits aconitase by forming a metal–nitrosyl complex with its iron–sulfur cluster [105]. Further, iron sequestration via DNIC formation may yield reduced iron bioavailability for iron–sulfur cluster assembly. In this context, *Arabidopsis* mutants impaired in the iron–sulfur cluster assembly pathway reveal DNA hypermethylation [106]. Moreover, hypermethylation could be a result of

enhanced DMT activity. In this context, increased DMT activity was observed in nuclear protein extracts treated with •NO [107].

4.3. Alteration in DNA Methylation Does Not Correlate with Gene Expression

Several recent studies indicated a weak association between differential DNA methylation and gene expression changes [108]. For instance, in *Arabidopsis* mutants impaired in the methylation cycle, *mat4* [61] and *ms1* [50], differential DNA methylation of genes was not associated with their expression. Consistent with these findings, differentially expressed genes displayed no significant differences in DNA methylation profiles between *gsnor1-3* and wt. Hence, these results indicate that transcriptional changes occur largely independently of detectable variation in the DNA methylation pattern. In this regard, only 4% of DMGs (genes overlapping with identified DMRs in their genic, 3kb up- and/or downstream region) were differentially expressed. This finding is comparable to previous studies. For instance, about 5% of DMGs were differentially expressed in *Arabidopsis* roots challenged with beet cyst nematode *Heterodera schachtii* [108]. Promotor methylation (3kb upstream region) was typically associated with gene repression; however, in some cases, it enhanced gene transcription in *gsnor1-3* (Table 3). Gene body methylation (between start and stop codons) seems to have a weak effect on gene expression in *Arabidopsis* [109,110], and its function remains enigmatic [111].

Nevertheless, constitutive mis-regulation of genes which are not directly targeted by DNA methylation may result from methylation-dependent alteration in the transcriptional networks [112]. The linkage between DEGs not targeted by differential DNA methylation and methylation-dependent alteration in the transcriptional network [62,112] is exemplified at the *PR1* gene. The *PR1* transcript is upregulated in mutants globally defective in the maintenance of CG (*met1*) or non-CG methylation (*ddc*) [112], whereas *PR1* is down-regulated in hypermethylated *35S::MS1* plants [62]. Likewise, *PR1* expression is reduced (Supplemental Table S7) and delayed [34] in *gsnor1-3*. Notably, mutants globally defective in DNA methylation were markedly resistant to *Pst* [112], whereas plants with an increased DNA methylation level (*35S::METS1*; *Arabidopsis* plants overexpressing *MS1*) and *gsnor1-3* showed attenuated resistance to *Pst* [34,62].

Besides altered DNA methylation levels, transcriptional changes are probably also caused by the pleiotropic effects of an impaired GSNOR1 function. For instance, loss of the GSNOR1 function caused the differential expression of several transcription factors (Supplemental Table S7). Further, proteins involved in transcriptional regulation were identified as targets for S-nitrosation [33]. Moreover, loss of the GSNOR1 function caused enhanced global levels of H3K27me2 (Table 1), which is usually highly enriched at the promoter of inactive genes [113]. Other reasons why loss of the GSNOR1 function induces transcriptional changes could be the modulation of the chromatin structure by other epigenetic mechanisms. For instance, non-coding miscellaneous RNAs are differentially expressed in response to GSNO [114]. In general, non-coding RNAs are regulators of gene expression by a variety of mechanisms such as chromatin remodeling, or they regulate gene expression at the transcriptional or post-transcriptional levels. Furthermore, transcriptional changes could be linked to the proximity of differentially methylated TEs to DEGs [108].

4.4. GSNOR1 Regulates Demethylation and Expression of TEs and Stress-Responsive Genes

GSNOR1 activity is required for the reduction in H3K9me2. H3K9me2 plays important roles in plant environmental stress response [115]. For instance, gene expression induced by ABA and salt stress is associated with the reduction in gene repression marks, such as H3K9me2, at ABA and abiotic stress-responsive genes [116]. In this context, lowering the H3K9me2 level at stress-related genes might be a regulatory mechanism of GSNOR1 to activate the stress response. Moreover, the repressive histone mark H3K9me2 is associated with TE silencing. Repression of TEs is required to guarantee genome stability. Therefore, TEs are generally located in transcriptionally silenced heterochromatic regions marked by DNA methylation and repressive histone modifications, such as H3K9me2 [43,115]. In *gsnor1-3*,

DNA methylation differs in the TE-rich pericentromeric region from wt (Figure 4). Indeed, parts are hyper- and hypomethylated. However, the genomic annotation of the identified DMRs resulted in mainly hypermethylated TEs (Figure 6A,D). Among them, LTR/Copia- and Line/L1-type TEs, predominantly regulated through H3K9me2 and non-CG DNA methylation pathways [117], but also LTR/Gypsy-type TEs, predominantly regulated by H3K27me1 methylation [117], were found. Consistent with the enhanced DNA methylation, the RNA-seq data indicate that TEs (expression analysis performed at family level) were mainly repressed in the *gsnor1-3* mutant (Figure 8A–C). The expression of transposons under plant stress, such as heat, cold, drought, wounding, viruses, and pathogens [118], is a well-known phenomenon [119–126]. According to McClintock [127], boosting the expression and transposition activity of TEs in environmental stress conditions results in extensive genomic re-structuring, which finally facilitates the adaptation of species and populations to a changing environment [128]. Moreover, the TEs closely associated with genic regions could be involved in directly reprogramming transcriptional networks, affecting the expression profiles of individual genes and fine-tuning the host response to specific stimuli [129,130]. In this context, the impaired plant disease responses [34,131] and the heat sensitivity [35] of GSNOR1-deficient *Arabidopsis* could, at least, be partly based on the reduced activation of TEs.

Interestingly, the GSNOR1 function is also required for the demethylation and expression of several stress-responsive genes, e.g., Flotillin-like protein1 and 2 (AT5G25250, AT5G25260), which are involved in the UV stress response, or cytochrome P450 94C1 (AT2G27690), which is involved in the wounding response (Table 3). Plant flotillins are a subgroup of the SPFH domain protein superfamily, consisting of three proteins, FLOT1, FLOT2, and FLOT3, in *A. thaliana*. *FLOT* genes respond differentially to different types of abiotic and biotic stresses, nutrient depletion, and phytohormones [132,133].

Cytochrome P450 94C1 encodes an enzyme involved in jasmonoyl-L-isoleucine (JA-Ile) oxidation. Jasmonic acid (JA) is an important signaling hormone exhibiting a broad spectrum of physiological activities in growth and development. JA also fulfills an important signaling function in plant defense, particularly the defense against insect herbivores and necrotrophic pathogens. In particular, the conjugate of jasmonate and isoleucine (JA-Ile) is a major regulator which controls gene expression and production of secondary metabolites after (a)biotic challenges. The two cytochromes P450 94B3 and 94C1 catalyze two successive oxidation steps of JA-Ile for catabolic turnover [134,135]. The oxidized derivatives of JA-Ile accumulate in wounded *Arabidopsis* leaves. CYP94C1 catalyzes the oxidation of 12OH-JA-Ile to 12C00H-JA-Ile, and its transcripts accumulate in response to stress and wounding [136]. However, plants overexpressing CYP94C1 display a strongly impaired defense gene induction as well as reduced disease resistance [135], suggesting that a coordinated turnover of JA-Ile is essential for an effective stress response. In this context, the reduced expression of CYP94C1 in *gsnor1-3* might be responsible for herbivory susceptibility, as demonstrated in GSNOR-silenced *Nicotiana attenuata* [137].

In conclusion, the GSNOR1 function is required for a controlled processing of the methylation cycle, for a reduction in the repressive H3K9me2 histone mark, and for TE activation to enable an effective stress response (Figure 9). These findings present a new function of •NO as an epigenetic regulator and provide a new insight into •NO signaling in plants.

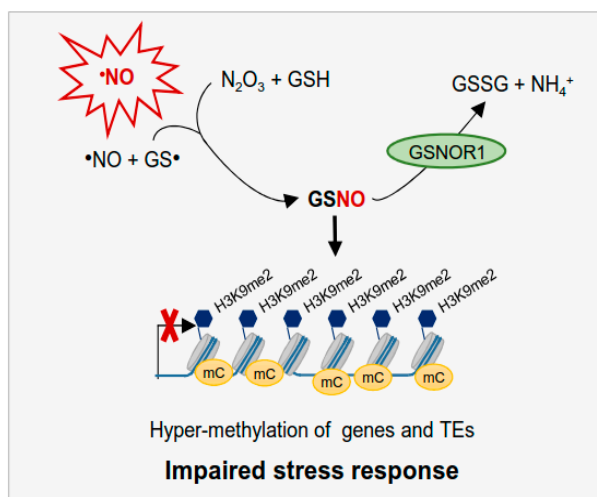


Figure 9. Proposed model illustrating the function of GSNOR1 in regulating methylation processes and expression of TEs and stress-responsive genes. •NO is endogenously produced under physiological conditions [18], and GSNO, as a more stable redox form of •NO, is formed and promotes methylation of H3K9 and DNA. Hypermethylation of TEs and stress-responsive genes results in impaired stress response. Since GSNO is enzymatically degraded by GSNOR1, GSNOR1 activity positively affects stress response by promoting expression of TEs and stress-responsive genes.

5. Conclusions

In this study, we demonstrated that the GSNOR1 function is required for SAM homeostasis, and, consequently, loss of GSNOR1 activity affects transmethylation reactions. We observed a significant global increase in the repressive H3K9me2 mark in *gsnor1-3*. H3K9me2-modified chromatin regions tightly correlate with methylated DNA regions. Whole-genome bisulfite sequencing and transcriptome analyses revealed enhanced DNA methylation and reduced expression of TEs and stress-responsive genes in *gsnor1-3*. This impaired expression of TEs and stress-responsive genes is in accordance with described susceptibility of *gsnor1-3* to e. g. pathogen infection and heat stress. In conclusion, our data suggest that GSNOR1 function is required to reduce the level of the repressive chromatin mark H3K9me2 and DNA methylation at distinct TEs and stress-responsive genes to enable effective stress response.

Supplementary Materials: The following are available online at <https://www.mdpi.com/article/10.3390/antiox10071128/s1>. Supplemental Table S1. Oligonucleotides used for the characterization of transgenic lines and cloning. Supplemental Table S2. DMRs identified in *gsnor1-3* and *sahh1* compared to wt. Supplemental Table S3. DMGs identified in *gsnor1-3* compared to wt. Supplemental Table S4. DMRs overlapping with TEs in *gsnor1-3*. Supplemental Table S5. DMGs identified in *sahh1* compared to wt. Supplemental Table S6. DMRs overlapping with TEs in *sahh1*. Supplemental Table S7. DEGs identified in *gsnor1-3* compared to wt. Supplemental Table S8. TE families differentially expressed in *gsnor1-3* compared to wt. Supplemental Table S9. DEGs identified in *sahh1* compared to wt. Supplemental Table S10. TE families differentially expressed in *sahh1* compared to wt. Supplemental Table S11. List of GO terms significantly enriched in the set of DEGs in *gsnor1-3*. Supplemental Table S12. List of GO terms significantly enriched in the set of DEGs in *sahh1*. **Supplemental Figure Legend:** Supplemental Figure S1. Loss of GSNOR1 function results in an increased RSNO content under basal conditions. Supplemental Figure S2. SAHH1 is S-nitrosated and inhibited by GSNO. Supplemental Figure S3. PCR-based genotyping of transgenic lines harboring *TS-GUS* insertion and *sahh1* or *gsnor1-3* mutation. Supplemental Figure S4. Annotation of DMRs to genomic features.

Supplemental Figure S5. DNA methylation is poorly correlated with gene expression differences in *gsnor1-3*. Supplemental Figure S6. DNA methylation is poorly correlated with gene expression differences in *sahh1*.

Author Contributions: Conceptualization, C.L.; formal analysis, E.R., P.H., I.F., E.G. and M.W.; investigation, E.R., P.H., I.F., Y.H. and M.W.; methodology, E.R., P.H. and M.W.; supervision, C.L.; writing—original draft, E.R.; writing—review and editing, R.H., A.I., C.B., J.D. and C.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Bundesministerium für Bildung und Forschung (BMBF). Research at Heidelberg, including the Metabolomics Core Technology Platform (MCTP), is supported by the German Research Foundation (grants: ZUK 49/2010–3009262, WI 3560/1-2, WI 3560/4-1, and HE 1848/15-2).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All the data analyzed for this manuscript are included. The analyzed data are available upon reasonable request to the corresponding author.

Acknowledgments: We thank Elke Mattes, Lucia Gößl, Rosina Ludwig, and Katharina Jandrasits for excellent technical assistance. This work was supported by the Bundesministerium für Bildung und Forschung (BMBF). Moreover, we thank the Metabolomics Core Technology Platform (MCTP) of the Excellence cluster “CellNetworks” (University of Heidelberg) for support with LC-based metabolite quantification. The Metabolomics Core Technology Platform (MCTP) is supported by the German Research Foundation (grant no. ZUK 49/2010–3009262, WI 3560/1-2, WI 3560/4-1, and HE 1848/15-2). We thank Hervé Vaucheret for providing seeds of the *TS-GUS* L5 transgenic *Arabidopsis* line, and Barbara Moffat for providing the anti-AtSAHH1 antibody.

Conflicts of Interest: The authors declare that they have no conflict of interest.

References

- Bethke, P.C.; Libourel, I.G.L.; Aoyama, N.; Chung, Y.-Y.; Still, D.W.; Jones, R.L. The Arabidopsis Aleurone Layer Responds to Nitric Oxide, Gibberellin, and Absciscic Acid and Is Sufficient and Necessary for Seed Dormancy. *Plant Physiol.* **2007**, *143*, 1173–1188. [\[CrossRef\]](#)
- Albertos, P.; Romero-Puertas, M.C.; Tatematsu, K.; Mateos, I.; Sanchez-Vicente, I.; Nambara, E.; Lorenzo, O. S-nitrosylation triggers ABI5 degradation to promote seed germination and seedling growth. *Nat. Commun.* **2015**, *6*, 8669. [\[CrossRef\]](#) [\[PubMed\]](#)
- Wang, P.; Zhu, J.K.; Lang, Z. Nitric oxide suppresses the inhibitory effect of abscisic acid on seed germination by S-nitrosylation of SnRK2 proteins. *Plant Signal. Behav.* **2015**, *10*, 2–5. [\[CrossRef\]](#) [\[PubMed\]](#)
- Correa-Aragunde, N.; Graziano, M.; Lamattina, L. Nitric oxide plays a central role in determining lateral root development in tomato. *Planta* **2004**, *218*, 900–905. [\[CrossRef\]](#) [\[PubMed\]](#)
- Zhao, D.Y.; Tian, Q.Y.; Li, L.H.; Zhang, W.H. Nitric oxide is involved in nitrate-induced inhibition of root elongation in *Zea mays*. *Ann. Bot.* **2007**, *100*, 497–503. [\[CrossRef\]](#) [\[PubMed\]](#)
- Beligni, M.V.; Lamattina, L. Nitric oxide stimulates seed germination and de-etiolation, and inhibits hypocotyl elongation, three light-inducible responses in plants. *Planta* **2000**, *210*, 215–221. [\[CrossRef\]](#)
- Laxalt, A.M.; García-Mata, C.; Lamattina, L. The Dual Role of Nitric Oxide in Guard Cells: Promoting and Attenuating the ABA and Phospholipid-Derived Signals Leading to the Stomatal Closure. *Front. Plant Sci.* **2016**, *7*, 2007–2010. [\[CrossRef\]](#) [\[PubMed\]](#)
- Hu, X.; Neill, S.J.; Tang, Z.; Cai, W. Nitric Oxide Mediates Gravitropic Bending in Soybean Roots. *Plant Physiol.* **2005**, *137*, 663–670. [\[CrossRef\]](#)
- He, Y.; Tang, R.-H.; Hao, Y.; Stevens, R.D.; Cook, C.W.; Ahn, S.M.; Jing, L.; Yang, Z.; Chen, L.; Guo, F.; et al. Nitric Oxide Represses the Arabidopsis Floral Transition. *Science* **2004**, *305*, 1968–1971. [\[CrossRef\]](#) [\[PubMed\]](#)
- Seligman, K.; Saviani, E.E.; Oliveira, H.C.; Pinto-Maglio, C.A.F.; Salgado, I. Floral transition and nitric oxide emission during flower development in *Arabidopsis thaliana* is affected in nitrate reductase-deficient plants. *Plant Cell Physiol.* **2008**, *49*, 1112–1121. [\[CrossRef\]](#)
- Prado, A.M. Nitric oxide is involved in growth regulation and re-orientation of pollen tubes. *Development* **2004**, *131*, 2707–2714. [\[CrossRef\]](#) [\[PubMed\]](#)
- Mishina, T.E.; Lamb, C.; Zeier, J. Expression of a nitric oxide degrading enzyme induces a senescence programme in Arabidopsis. *Plant Cell Environ.* **2007**, *30*, 39–52. [\[CrossRef\]](#)
- Trapet, P.; Kulik, A.; Lamotte, O.; Jeandroz, S.; Bourque, S.; Nicolas-Francis, V.; Rosnoblet, C.; Besson-Bard, A.; Wendehenne, D. NO signaling in plant immunity: A tale of messengers. *Phytochemistry* **2015**, *112*, 72–79. [\[CrossRef\]](#) [\[PubMed\]](#)

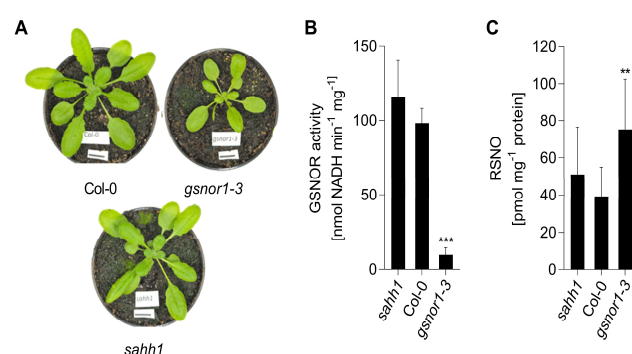
14. Yu, M.; Lamattina, L.; Spoel, S.H.; Loake, G.J. Nitric oxide function in plant biology: A redox cue in deconvolution. *New Phytol.* **2014**, *202*, 1142–1156. [\[CrossRef\]](#)
15. Fancy, N.N.; Bahlmann, A.-K.; Loake, G.J. Nitric oxide function in plant abiotic stress. *Plant. Cell Environ.* **2017**, *40*, 462–472. [\[CrossRef\]](#) [\[PubMed\]](#)
16. Buet, A.; Simontacchi, M. Nitric oxide and plant iron homeostasis. *Ann. N. Y. Acad. Sci.* **2015**, *1340*, 39–46. [\[CrossRef\]](#)
17. Gross, F.; Durner, J.; Gaupels, F. Nitric oxide, antioxidants and prooxidants in plant defence responses. *Front Plant Sci* **2013**, *4*, 419. [\[CrossRef\]](#) [\[PubMed\]](#)
18. Astier, J.; Gross, I.; Durner, J. Nitric oxide production in plants: An update. *J. Exp. Bot.* **2018**, *69*, 3401–3411. [\[CrossRef\]](#)
19. Wurm, C.J.; Lindermayr, C. Nitric oxide signaling in the plant nucleus: The function of nitric oxide in chromatin modulation and transcription. *J. Exp. Bot.* **2021**, *72*, 808–818. [\[CrossRef\]](#)
20. Kolbert, Z.; Feigl, G.; Bordé, Á.; Molnár, Á.; Erdei, L. Protein tyrosine nitration in plants: Present knowledge, computational prediction and future perspectives. *Plant Physiol. Biochem.* **2017**, *113*, 56–63. [\[CrossRef\]](#)
21. Kovacs, I.; Lindermayr, C. Nitric oxide-based protein modification: Formation and site-specificity of protein S-nitrosylation. *Front. Plant Sci* **2013**, *4*, 137. [\[CrossRef\]](#)
22. Feng, J.; Chen, L.; Zuo, J. Protein S-Nitrosylation in plants: Current progresses and challenges. *J. Integr. Plant Biol.* **2019**, *61*, 1206–1223. [\[CrossRef\]](#) [\[PubMed\]](#)
23. Ageeva-Kieferle, A.; Rudolf, E.E.; Lindermayr, C. Redox-Dependent Chromatin Remodeling: A New Function of Nitric Oxide as Architect of Chromatin Structure in Plants. *Front. Plant Sci.* **2019**, *10*, 625. [\[CrossRef\]](#) [\[PubMed\]](#)
24. Kovacs, I.; Ageeva, A.; König, E.-E.; Lindermayr, C. Chapter Two—S-Nitrosylation of Nuclear Proteins: New Pathways in Regulation of Gene Expression. In *Nitric Oxide and Signaling in Plants*; Wendehenne, D., Ed.; Academic Press: London, UK, 2016; Volume 77, pp. 15–39.
25. Mengel, A.; Chaki, M.; Shekariesfahlan, A.; Lindermayr, C. Effect of nitric oxide on gene transcription—S-nitrosylation of nuclear proteins. *Front. Plant Sci.* **2013**, *4*, 293. [\[CrossRef\]](#) [\[PubMed\]](#)
26. León, J.; Costa, Á.; Castillo, M.C. Nitric oxide triggers a transient metabolic reprogramming in Arabidopsis. *Sci. Rep.* **2016**, *6*, 37945. [\[CrossRef\]](#)
27. Freschi, L. Nitric oxide and phytohormone interactions: Current status and perspectives. *Front Plant Sci.* **2013**, *4*, 398. [\[CrossRef\]](#)
28. Simontacchi, M.; García-Mata, C.; Bartoli, C.G.; Santa-Maria, G.E.; Lamattina, L. Nitric oxide as a key component in hormone-regulated processes. *Plant Cell Rep.* **2013**, *32*, 853–866. [\[CrossRef\]](#)
29. Lindermayr, C. Crosstalk between reactive oxygen species and nitric oxide in plants: Key role of S-nitrosogluthathione reductase. *Free Radic. Biol. Med.* **2018**, *122*, 110–115. [\[CrossRef\]](#)
30. Jahnová, J.; Luhová, L.; Petřivalský, M. S-Nitrosogluthathione Reductase—The Master Regulator of Protein S-Nitrosylation in Plant NO Signaling. *Plants* **2019**, *8*, 48. [\[CrossRef\]](#)
31. Begara-Morales, J.C.; Chaki, M.; Valderrama, R.; Sánchez-Calvo, B.; Mata-Pérez, C.; Padilla, M.N.; Corpas, F.J.; Barroso, J.B. Nitric oxide buffering and conditional nitric oxide release in stress response. *J. Exp. Bot.* **2018**, *69*, ery072. [\[CrossRef\]](#)
32. Floryszak-Wieczorek, J.; Milczarek, G.; Arasimowicz, M.; Ciszewski, A. Do nitric oxide donors mimic endogenous NO-related response in plants? *Planta* **2006**, *224*, 1363–1372. [\[CrossRef\]](#) [\[PubMed\]](#)
33. Hu, J.; Huang, X.; Chen, L.; Sun, X.; Lu, C.; Zhang, L.; Wang, Y.; Zuo, J. Site-specific nitrosoproteomic identification of endogenously S-nitrosylated proteins in Arabidopsis. *Plant Physiol.* **2015**, *167*, 1731–1746. [\[CrossRef\]](#)
34. Feechan, A.; Kwon, E.; Yun, B.W.; Wang, Y.; Pallas, J.A.; Loake, G.J. A central role for S-nitrosothiols in plant disease resistance. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 8054–8059. [\[CrossRef\]](#) [\[PubMed\]](#)
35. Lee, U.; Wie, C.; Fernandez, B.O.; Feelisch, M.; Vierling, E. Modulation of nitrosative stress by S-nitrosogluthathione reductase is critical for thermotolerance and plant growth in Arabidopsis. *Plant Cell* **2008**, *20*, 786–802. [\[CrossRef\]](#) [\[PubMed\]](#)
36. Chen, R.; Sun, S.; Wang, C.; Li, Y.; Liang, Y.; An, F.; Li, C.; Dong, H.; Yang, X.; Zhang, J.; et al. The Arabidopsis PARAQUAT RESISTANT2 gene encodes an S-nitrosogluthathione reductase that is a key regulator of cell death. *Cell Res.* **2009**, *19*, 1377–1387. [\[CrossRef\]](#)
37. Kawabe, H.; Ohtani, M.; Kurata, T.; Sakamoto, T.; Demura, T. Protein S-Nitrosylation Regulates Xylem Vessel Cell Differentiation in Arabidopsis. *Plant Cell Physiol.* **2018**, *59*, 17–29. [\[CrossRef\]](#)
38. Yun, B.W.; Skelly, M.J.; Yin, M.; Yu, M.; Mun, B.G.; Lee, S.U.; Hussain, A.; Spoel, S.H.; Loake, G.J. Nitric oxide and S-nitrosogluthathione function additively during plant immunity. *New Phytol.* **2016**, *211*, 516–526. [\[CrossRef\]](#) [\[PubMed\]](#)
39. Kovacs, I.; Holzmeister, C.; Wirtz, M.; Geerlof, A.; Fröhlich, T.; Römmling, G.; Kuruthukulangarakoola, G.T.; Linster, E.; Hell, R.; Arnold, G.J.; et al. ROS-Mediated Inhibition of S-nitrosogluthathione Reductase Contributes to the Activation of Anti-oxidative Mechanisms. *Front. Plant Sci.* **2016**, *7*, 1669. [\[CrossRef\]](#)
40. Chaki, M.; Shekariesfahlan, A.; Ageeva, A.; Mengel, A.; von Toerne, C.; Durner, J.; Lindermayr, C. Identification of nuclear target proteins for S-nitrosylation in pathogen-treated Arabidopsis thaliana cell cultures. *Plant Sci.* **2015**, *238*, 115–126. [\[CrossRef\]](#)
41. Imran, Q.M.; Hussain, A.; Lee, S.U.; Mun, B.G.; Falak, N.; Loake, G.J.; Yun, B.W. Transcriptome profile of NO-induced Arabidopsis transcription factor genes suggests their putative regulatory role in multiple biological processes. *Sci. Rep.* **2018**, *8*, 771. [\[CrossRef\]](#)
42. Schvartzman, J.M.; Thompson, C.B.; Finley, L.W.S. Metabolic regulation of chromatin modifications and gene expression. *J. Cell Biol.* **2018**, *217*, 2247–2259. [\[CrossRef\]](#)

43. Zhang, H.; Lang, Z.; Zhu, J.K. Dynamics and function of DNA methylation in plants. *Nat. Rev. Mol. Cell Biol.* **2018**, *19*, 489–506. [\[CrossRef\]](#)
44. Mengel, A.; Ageeva, A.; Georgii, E.; Bernhardt, J.; Wu, K.; Durner, J.; Lindermayr, C. Nitric Oxide Modulates Histone Acetylation at Stress Genes by Inhibition of Histone Deacetylases. *Plant Physiol.* **2017**, *173*, 1434–1452. [\[CrossRef\]](#)
45. Li, H.; Torres-Garcia, J.; Latrasse, D.; Benhamed, M.; Schilderink, S.; Zhou, W.; Kulikova, O.; Hirt, H.; Bisseling, T. Plant-specific histone deacetylases HDT1/2 regulate GIBBERELLIN 2-OXIDASE2 expression to control arabidopsis root meristem cell number. *Plant Cell* **2017**, *29*, 2183–2196. [\[CrossRef\]](#) [\[PubMed\]](#)
46. Ou, X.; Zhuang, T.; Yin, W.; Miao, Y.; Wang, B.; Zhang, Y.; Lin, X.; Xu, C.; von Wettstein, D.; Rustgi, S.; et al. DNA Methylation Changes Induced in Rice by Exposure to High Concentrations of the Nitric Oxide Modulator, Sodium Nitroprusside. *Plant Mol. Biol. Report.* **2015**, *33*, 1428–1440. [\[CrossRef\]](#)
47. Su, X.; Wellen, K.E.; Rabinowitz, J.D. Metabolic control of methylation and acetylation. *Curr. Opin. Chem. Biol.* **2016**, *30*, 52–60. [\[CrossRef\]](#) [\[PubMed\]](#)
48. Gorelova, V.; Ambach, L.; Rébeillé, F.; Stove, C.; Van Der Straeten, D. Folate in Plants: Research Advances and Progress in Crop Biofortification. *Front. Chem.* **2017**, *5*, 21. [\[CrossRef\]](#)
49. Sauter, M.; Moffatt, B.; Saechao, M.C.; Hell, R.; Wirtz, M. Methionine salvage and S-adenosylmethionine: Essential links between sulfur, ethylene and polyamine biosynthesis. *Biochem. J.* **2013**, *451*, 145–154. [\[CrossRef\]](#) [\[PubMed\]](#)
50. Yan, X.; Ma, L.; Pang, H.; Wang, P.; Liu, L.; Cheng, Y.; Cheng, J.; Guo, Y.; Li, Q. METHIONINE SYNTHASE1 is involved in chromatin silencing by maintaining dna and histone methylation. *Plant Physiol.* **2019**, *181*, 249–261. [\[CrossRef\]](#)
51. Huang, X.-Y.; Li, M.; Luo, R.; Zhao, F.-J.; Salt, D.E. Epigenetic regulation of sulfur homeostasis in plants. *J. Exp. Bot.* **2019**, *70*, 4171–4182. [\[CrossRef\]](#) [\[PubMed\]](#)
52. Rocha, P.S.; Sheikh, M.; Melchiorre, R.; Fagard, M.; Boutet, S.; Loach, R.; Moffatt, B.; Wagner, C.; Vaucheret, H.; Furner, I. The Arabidopsis HOMOLOG-DEPENDENT GENE SILENCING1 gene codes for an S-adenosyl-L-homocysteine hydrolase required for DNA methylation-dependent gene silencing. *Plant Cell* **2005**, *17*, 404–417. [\[CrossRef\]](#)
53. Baubec, T.; Dinh, H.Q.; Pecinka, A.; Rakic, B.; Rozhon, W.; Wohlrab, B.; von Haeseler, A.; Scheid, O.M. Cooperation of multiple chromatin modifications can generate unanticipated stability of epigenetic States in Arabidopsis. *Plant Cell* **2010**, *22*, 34–47. [\[CrossRef\]](#)
54. Li, C.H.; Yu, N.; Jiang, S.M.; Shanguan, X.X.; Wang, L.J.; Chen, X.Y. Down-regulation of S-adenosyl-L-homocysteine hydrolase reveals a role of cytokinin in promoting transmethylation reactions. *Planta* **2008**, *228*, 125–136. [\[CrossRef\]](#)
55. Jordan, N.D.; West, J.P.; Bottley, A.; Sheikh, M.; Furner, I. Transcript profiling of the hypomethylated hog1 mutant of Arabidopsis. *Plant Mol. Biol.* **2007**, *65*, 571–586. [\[CrossRef\]](#)
56. Ouyang, B.; Fei, Z.; Joung, J.G.; Kolenovsky, A.; Koh, C.; Nowak, J.; Caplan, A.; Keller, W.A.; Cui, Y.; Cutler, A.J.; et al. Transcriptome profiling and methyl homeostasis of an Arabidopsis mutant deficient in S-adenosylhomocysteine hydrolase1 (SAHH1). *Plant Mol. Biol.* **2012**, *79*, 315–331. [\[CrossRef\]](#) [\[PubMed\]](#)
57. Tanaka, H.; Masuta, C.; Uehara, K.; Kataoka, J.; Koiwai, A.; Noma, M. Morphological changes and hypomethylation of DNA in transgenic tobacco expressing antisense RNA of the S-adenosyl-L-homocysteine hydrolase gene. *Plant Mol. Biol.* **1997**, *35*, 981–986. [\[CrossRef\]](#)
58. Fojtová, M.; Kovarik, A.; Votruba, I.; Holý, A. Evaluation of the impact of S-adenosylhomocysteine metabolic pools on cytosine methylation of the tobacco genome. *Eur. J. Biochem.* **1998**, *252*, 347–352. [\[CrossRef\]](#) [\[PubMed\]](#)
59. Kovavik, A.; Koukalová, B.; Holý, A.; Bezděk, M. Sequence-specific hypomethylation of the tobacco genome induced with dihydroxypropyladenine, ethionine and 5-azacytidine. *FEBS Lett.* **1994**, *353*, 309–311. [\[CrossRef\]](#)
60. Fulnecek, J.; Matyasek, R.; Votruba, I.; Holy, A.; Krizova, K.; Kovarik, A. Inhibition of SAH-hydrolase activity during seed germination leads to deregulation of flowering genes and altered flower morphology in tobacco. *Mol. Genet. Genom.* **2011**, *285*, 225–236. [\[CrossRef\]](#) [\[PubMed\]](#)
61. Meng, J.; Wang, L.; Wang, J.; Zhao, X.; Cheng, J.; Yu, W.; Jin, D.; Li, Q.; Gong, Z. METHIONINE ADENOSYLTRANSFERASE 4 mediates DNA and histone methylation. *Plant Physiol.* **2018**, *177*, 652–670. [\[CrossRef\]](#) [\[PubMed\]](#)
62. González, B.; Vera, P. Folate Metabolism Interferes with Plant Immunity through 1C Methionine Synthase-Directed Genome-wide DNA Methylation Enhancement. *Mol. Plant* **2019**, *12*, 1227–1242. [\[CrossRef\]](#)
63. Wu, X.; Li, F.; Kolenovsky, A.; Caplan, A.; Cui, Y.; Cutler, A.; Tsang, E.W.T. A mutant deficient in S-adenosylhomocysteine hydrolase in Arabidopsis shows defects in root-hair development. *Botany* **2009**, *87*, 571–584. [\[CrossRef\]](#)
64. Morel, J.-B.; Mourrain, P.; Béclin, C.; Vaucheret, H. DNA methylation and chromatin structure affect transcriptional and post-transcriptional transgene silencing in Arabidopsis. *Curr. Biol.* **2000**, *10*, 1591–1594. [\[CrossRef\]](#)
65. Murashige, T.; Skoog, F. A Revised Medium for Rapid Growth and Bio Assays with Tobacco Tissue Cultures. *Physiol. Plant.* **1962**, *15*, 473–497. [\[CrossRef\]](#)
66. Bürstenbinder, K.; Rzewuski, G.; Wirtz, M.; Hell, R.; Sauter, M. The role of methionine recycling for ethylene synthesis in Arabidopsis. *Plant J.* **2007**, *49*, 238–249. [\[CrossRef\]](#)
67. Krueger, F.; Andrews, S.R. Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **2011**, *27*, 1571–1572. [\[CrossRef\]](#)
68. Langmead, B.; Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **2012**, *9*, 357–359. [\[CrossRef\]](#)

69. Becker, C.; Hagmann, J.; Müller, J.; Koenig, D.; Stegle, O.; Borgwardt, K.; Weigel, D. Spontaneous epigenetic variation in the *Arabidopsis thaliana* methylome. *Nature* **2011**, *480*, 245–249. [\[CrossRef\]](#)
70. Hagmann, J.; Becker, C.; Müller, J.; Stegle, O.; Meyer, R.C.; Wang, G.; Schneeberger, K.; Fitz, J.; Altmann, T.; Bergelson, J.; et al. Century-scale Methylome Stability in a Recently Diverged *Arabidopsis thaliana* Lineage. *PLoS Genet.* **2015**, *11*, e1004920. [\[CrossRef\]](#) [\[PubMed\]](#)
71. Richards, C.L.; Alonso, C.; Becker, C.; Bossdorf, O.; Bucher, E.; Colomé-Tatché, M.; Durka, W.; Engelhardt, J.; Gaspar, B.; Gogol-Döring, A.; et al. Ecological plant epigenetics: Evidence from model and non-model species, and the way forward. *Ecol. Lett.* **2017**, *20*, 1576–1590. [\[CrossRef\]](#) [\[PubMed\]](#)
72. Dobin, A.; Davis, C.A.; Schlesinger, F.; Drenkow, J.; Zaleski, C.; Jha, S.; Batut, P.; Chaisson, M.; Gingeras, T.R. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **2013**, *29*, 15–21. [\[CrossRef\]](#)
73. Bray, N.L.; Pimentel, H.; Melsted, P.; Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **2016**, *34*, 525–527. [\[CrossRef\]](#)
74. Love, M.I.; Huber, W.; Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **2014**, *15*, 550. [\[CrossRef\]](#)
75. Cheng, Y.T.; Germain, H.; Wiermer, M.; Bi, D.; Xu, F.; Garcia, A.V.; Wirthmueller, L.; Despres, C.; Parker, J.E.; Zhang, Y.; et al. Nuclear pore complex component MOS7/Nup88 is required for innate immunity and nuclear accumulation of defense regulators in *Arabidopsis*. *Plant Cell* **2009**, *21*, 2503–2516. [\[CrossRef\]](#) [\[PubMed\]](#)
76. Feller, C.; Forné, I.; Imhof, A.; Becker, P.B. Global and specific responses of the histone acetylome to systematic perturbation. *Mol. Cell* **2015**, *57*, 559–572. [\[CrossRef\]](#)
77. Peri, S.; Steen, H.; Pandey, A. GPMW—A software tool for analyzing proteins and peptides. *Trends Biochem. Sci.* **2001**, *26*, 687–689. [\[CrossRef\]](#)
78. Lindermayr, C.; Saalbach, G.; Durner, J. Proteomic identification of S-nitrosylated proteins in *Arabidopsis*. *Plant Physiol.* **2005**, *137*, 921–930. [\[CrossRef\]](#) [\[PubMed\]](#)
79. Fares, A.; Rossignol, M.; Peltier, J.B. Proteomics investigation of endogenous S-nitrosylation in *Arabidopsis*. *Biochem. Biophys. Res. Commun.* **2011**, *416*, 331–336. [\[CrossRef\]](#)
80. Maldonado-Alconada, A.M.; Echevarría-Zomeño, S.; Lindermayr, C.; Redondo-López, I.; Durner, J.; Jorrín-Novo, J.V. Proteomic analysis of *Arabidopsis* protein S-nitrosylation in response to inoculation with *Pseudomonas syringae*. *Acta Physiol. Plant.* **2011**, *33*, 1493–1514. [\[CrossRef\]](#)
81. Puyaubert, J.; Fares, A.; Reze, N.; Peltier, J.B.; Baudouin, E.; Rézé, N.; Peltier, J.B.; Baudouin, E.; Reze, N.; Peltier, J.B.; et al. Identification of endogenously S-nitrosylated proteins in *Arabidopsis* plantlets: Effect of cold stress on cysteine nitrosylation level. *Plant Sci.* **2014**, *215*, 150–156. [\[CrossRef\]](#) [\[PubMed\]](#)
82. Puyaubert, J.; Baudouin, E. New clues for a cold case: Nitric oxide response to low temperature. *Plant Cell Env.* **2014**, *37*, 2623–2630. [\[CrossRef\]](#)
83. Song, X.; Cao, X. Context and Complexity: Analyzing Methylation in Trinucleotide Sequences. *Trends Plant Sci.* **2017**, *22*, 351–353. [\[CrossRef\]](#)
84. Du, J.; Johnson, L.M.; Jacobsen, S.E.; Patel, D.J. DNA methylation pathways and their crosstalk with histone methylation. *Nat. Rev. Mol. Cell Biol.* **2015**, *16*, 519–532. [\[CrossRef\]](#) [\[PubMed\]](#)
85. Cokus, S.J.; Feng, S.; Chen, Z.; Merriman, B.; Haudenschild, C.D.; Pradhan, S.; Nelson, S.F.; Pellegrini, M.; Jacobsen, S.E. Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. *Nature* **2008**, *452*, 215–219. [\[CrossRef\]](#) [\[PubMed\]](#)
86. Huang, X.; Zhang, S.; Li, K.; Thimmapuram, J.; Xie, S. ViewBS: A powerful toolkit for visualization of high-throughput bisulfite sequencing data. *Bioinformatics* **2018**, *34*, 708–709. [\[CrossRef\]](#) [\[PubMed\]](#)
87. Hoffman, D.R.; Cornatzer, W.E.; Duerre, J.A. Relationship between tissue levels of S-adenosylmethionine, S-adenosylhomocysteine, and transmethylation reactions. *Can. J. Biochem.* **1979**, *57*, 56–64. [\[CrossRef\]](#)
88. García-Giménez, J.L.; Romá-Mateo, C.; Pérez-Machado, G.; Peiró-Chova, L.; Pallardó, F.V. Role of glutathione in the regulation of epigenetic mechanisms in disease. *Free Radic. Biol. Med.* **2017**, *112*, 36–48. [\[CrossRef\]](#)
89. Detich, N.; Hamm, S.; Just, G.; Knox, J.D.; Szyf, M. The methyl donor S-Adenosylmethionine inhibits active demethylation of DNA. A candidate novel mechanism for the pharmacological effects of S-Adenosylmethionine. *J. Biol. Chem.* **2003**, *278*, 20812–20820. [\[CrossRef\]](#) [\[PubMed\]](#)
90. Pascale, R.M.; Simile, M.M.; Satta, G.; Seddaiu, M.A.; Daino, L.; Pinna, G.; Gaspa, M.A.L.V.; Feo, F. Comparative effects of L-methionine, S-adenosyl-L-methionine and 5'-methylthioadenosine on the growth of preneoplastic lesions and DNA methylation in rat liver during the early stages of hepatocarcinogenesis. *Anticancer Res.* **1991**, *11*, 1617–1624.
91. Ara, A.I.; Xia, M.; Ramani, K.; Mato, J.M.; Lu, S.C. S-Adenosylmethionine Inhibits Lipopolysaccharide-Induced Gene Expression via Modulation of Histone Methylation. *Hepatology* **2008**, *47*, 1655–1666. [\[CrossRef\]](#)
92. Speckmann, B.; Schulz, S.; Hiller, F.; Hesse, D.; Schumacher, F.; Kleuser, B.; Geisel, J.; Obeid, R.; Grune, T.; Kipp, A.P. Selenium increases hepatic DNA methylation and modulates one-carbon metabolism in the liver of mice. *J. Nutr. Biochem.* **2017**, *48*, 112–119. [\[CrossRef\]](#)
93. Luka, Z.; Capdevila, A.; Mato, J.M.; Wagner, C. A Glycine N-methyltransferase knockout mouse model for humans with deficiency of this enzyme. *Transgenic Res.* **2006**, *15*, 393–397. [\[CrossRef\]](#)

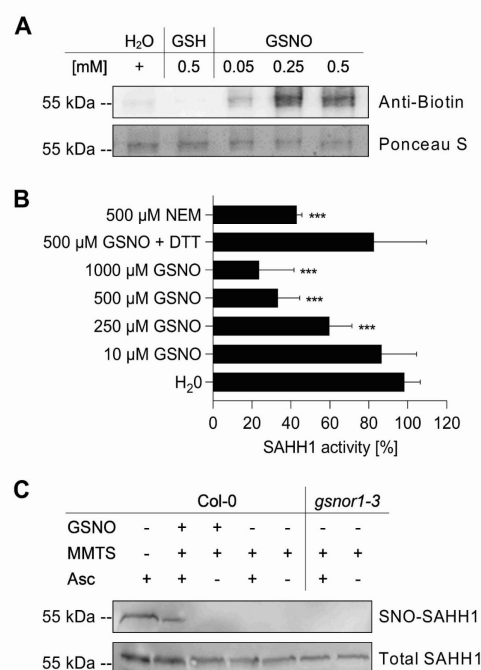
94. Lindermayr, C.; Rudolf, E.E.; Durner, J.; Groth, M. Interactions between metabolism and chromatin in plant models. *Mol. Metab.* **2020**, *38*, 100951. [\[CrossRef\]](#)
95. Hu, J.; Yang, H.; Mu, J.; Lu, T.; Peng, J.; Deng, X.; Kong, Z.; Bao, S.; Cao, X.; Zuo, J. Nitric Oxide Regulates Protein Methylation during Stress Responses in Plants. *Mol. Cell* **2017**, *67*, 702–710. [\[CrossRef\]](#)
96. Li, X.; Harris, C.J.; Zhong, Z.; Chen, W.; Liu, R.; Jia, B.; Wang, Z.; Li, S.; Jacobsen, S.E.; Du, J. Mechanistic insights into plant SUVH family H3K9 methyltransferases and their binding to context-biased non-CG DNA methylation. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, E8793–E8802. [\[CrossRef\]](#) [\[PubMed\]](#)
97. Wierzbicki, A.T.; Haag, J.R.; Pikaard, C.S. Noncoding transcription by RNA polymerase Pol IVb/Pol V mediates transcriptional silencing of overlapping and adjacent genes. *Cell* **2008**, *135*, 635–648. [\[CrossRef\]](#) [\[PubMed\]](#)
98. Zheng, B.; Wang, Z.; Li, S.; Yu, B.; Liu, J.Y.; Chen, X. Intergenic transcription by RNA polymerase II coordinates Pol IV and Pol V in siRNA-directed transcriptional gene silencing in Arabidopsis. *Genes Dev.* **2009**, *23*, 2850–2860. [\[CrossRef\]](#) [\[PubMed\]](#)
99. Shin, J.H.; Wang, H.L.V.; Lee, J.; Dinwiddie, B.L.; Belostotsky, D.A.; Chekanova, J.A. The Role of the Arabidopsis Exosome in siRNA-Independent Silencing of Heterochromatic Loci. *PLoS Genet.* **2013**, *9*, e1003411. [\[CrossRef\]](#) [\[PubMed\]](#)
100. Liu, Z.W.; Shao, C.R.; Zhang, C.J.; Zhou, J.X.; Zhang, S.W.; Li, L.; Chen, S.; Huang, H.W.; Cai, T.; He, X.J. The SET Domain Proteins SUVH2 and SUVH9 Are Required for Pol V Occupancy at RNA-Directed DNA Methylation Loci. *PLoS Genet.* **2014**, *10*, e1003948. [\[CrossRef\]](#) [\[PubMed\]](#)
101. Szyf, M. Epigenetics, DNA methylation, and chromatin modifying drugs. *Annu. Rev. Pharmacol. Toxicol.* **2009**, *49*, 243–263. [\[CrossRef\]](#)
102. Bovee, R.; Pham, V.; Fernandez, J.; Tretyakova, N.; Thomas, D.D. Nitric oxide is an epigenetic regulator of gene expression by directly controlling DNA methylation patterns. *Free Radic. Biol. Med.* **2018**, *120*, S114. [\[CrossRef\]](#)
103. Li, Y.; Kumar, S.; Qian, W. Active DNA demethylation: Mechanism and role in plant development. *Plant Cell Rep.* **2018**, *37*, 77–85. [\[CrossRef\]](#)
104. Crack, J.C.; Green, J.; Thomson, A.J.; Brun, N.E.L. Iron-sulfur clusters as biological sensors: The chemistry of reactions with molecular oxygen and nitric oxide. *Acc. Chem. Res.* **2014**, *47*, 3196–3205. [\[CrossRef\]](#)
105. Navarre, D.A.; Wendehenne, D.; Durner, J.; Noad, R.; Klessig, D.F. Nitric oxide modulates the activity of tobacco aconitase. *Plant Physiol.* **2000**, *122*, 573–582. [\[CrossRef\]](#)
106. Wang, X.; Li, Q.; Yuan, W.; Cao, Z.; Qi, B.; Kumar, S.; Li, Y.; Qian, W. The cytosolic Fe-S cluster assembly component MET18 is required for the full enzymatic activity of ROS1 in active DNA demethylation. *Sci. Rep.* **2016**, *6*, 26443. [\[CrossRef\]](#)
107. Hmadcha, B.A.; Bedoya, F.J.; Sobrino, F.; Pintado, E. Methylation-dependent Gene Silencing Induced by Interleukin 1 β Via Nitric Oxide Production. *J. Exp. Med.* **1999**, *190*, 1595–1603. [\[CrossRef\]](#)
108. Hewezi, T.; Lane, T.; Piya, S.; Rambani, A.; Rice, J.H.; Staton, M. Cyst Nematode Parasitism Induces Dynamic Changes in the Root Epigenome. *Plant Physiol.* **2017**, *174*, 405–420. [\[CrossRef\]](#)
109. Zhang, X.; Yazaki, J.; Sundaresan, A.; Cokus, S.; Chan, S.W.; Chen, H.; Henderson, I.R.; Shinn, P.; Pellegrini, M.; Jacobsen, S.E.; et al. Genome-wide high-resolution mapping and functional analysis of DNA methylation in Arabidopsis. *Cell* **2006**, *126*, 1189–1201. [\[CrossRef\]](#)
110. Takuno, S.; Gaut, B.S. Gene body methylation is conserved between plant orthologs and is of evolutionary consequence. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 1797–1802. [\[CrossRef\]](#)
111. Bewick, A.J.; Schmitz, R.J. Gene body DNA methylation in plants. *Curr. Opin. Plant Biol.* **2017**, *36*, 103–110. [\[CrossRef\]](#)
112. Downen, R.H.; Pelizzola, M.; Schmitz, R.J.; Lister, R.; Downen, J.M.; Nery, J.R.; Dixon, J.E.; Ecker, J.R. Widespread dynamic DNA methylation in response to biotic stress. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, E2183–E2191. [\[CrossRef\]](#) [\[PubMed\]](#)
113. Xiao, J.; Lee, U.S.; Wagner, D. Tug of war: Adding and removing histone lysine methylation in Arabidopsis. *Curr. Opin. Plant Biol.* **2016**, *34*, 41–53. [\[CrossRef\]](#)
114. Begara-Morales, J.C.; Sánchez-Calvo, B.; Luque, F.; Leyva-Pérez, M.O.; Leterrier, M.; Corpas, F.J.; Barroso, J.B. Differential transcriptomic analysis by RNA-seq of GSNOR-responsive genes between Arabidopsis roots and leaves. *Plant Cell Physiol.* **2014**, *55*, 1080–1095. [\[CrossRef\]](#)
115. Xu, L.; Jiang, H. Writing and Reading Histone H3 Lysine 9 Methylation in Arabidopsis. *Front. Plant Sci.* **2020**, *11*, 452. [\[CrossRef\]](#)
116. Chen, L.T.; Wu, K. Role of histone deacetylases HDA6 and HDA19 in ABA and abiotic stress response. *Plant Signal. Behav.* **2010**, *5*, 1318–1320. [\[CrossRef\]](#)
117. Stroud, H.; Hale, C.J.; Feng, S.; Caro, E.; Jacob, Y.; Michaels, S.D.; Jacobsen, S.E. DNA methyltransferases are required to induce heterochromatic re-replication in Arabidopsis. *PLoS Genet.* **2012**, *8*, e1002808. [\[CrossRef\]](#)
118. Negi, P.; Rai, A.N.; Suprasanna, P. Moving through the stressed genome: Emerging regulatory roles for transposons in plant stress response. *Front. Plant Sci.* **2016**, *7*, 1448. [\[CrossRef\]](#)
119. Bouvet, G.F.; Jacobi, V.; Plourde, K.V.; Bernier, L. Stress-induced mobility of OPHIO1 and OPHIO2, DNA transposons of the Dutch elm disease fungi. *Fungal Genet. Biol.* **2008**, *45*, 565–578. [\[CrossRef\]](#) [\[PubMed\]](#)
120. Johns, M.A.; Mottinger, J.; Freeling, M. A low copy number, copia-like transposon in maize. *EMBO J.* **1985**, *4*, 1093–1101. [\[CrossRef\]](#)
121. Hirochika, H. Activation of tobacco retrotransposons during tissue culture. *EMBO J.* **1993**, *12*, 2521–2528. [\[CrossRef\]](#) [\[PubMed\]](#)
122. Hirochika, H.; Sugimoto, K.; Otsuki, Y.; Tsugawa, H.; Kanda, M. Retrotransposons of rice involved in mutations induced by tissue culture. *Proc. Natl. Acad. Sci. USA* **1996**, *93*, 7783–7788. [\[CrossRef\]](#)

123. Turcich, M.P.; Bokhari-Riza, A.; Hamilton, D.A.; He, C.; Messier, W.; Stewart, C.-B.; Mascarenhas, J.P. PREM-2, a copia-type retroelement in maize is expressed preferentially in early microspores. *Sex. Plant Reprod.* **1996**, *9*, 65–74. [\[CrossRef\]](#)
124. Grandbastien, M.-A.; Lucas, H.; Morel, J.-B.; Mhiri, C.; Vernhettes, S.; Casacuberta, J.M. The expression of the tobacco Tnt1 retrotransposon is linked to plant defense responses. *Genetica* **1997**, *100*, 241–252. [\[CrossRef\]](#)
125. Grandbastien, M.-A.; Audeon, C.; Bonnivard, E.; Casacuberta, J.M.; Chalhoub, B.; Costa, A.-P.P.; Le, Q.H.; Melayah, D.; Petit, M.; Poncet, C.; et al. Stress activation and genomic impact of Tnt1 retrotransposons in Solanaceae. *Cytogenet. Genome Res.* **2005**, *110*, 229–241. [\[CrossRef\]](#)
126. Bui, Q.T.; Grandbastien, M.-A. LTR Retrotransposons as Controlling Elements of Genome Response to Stress? In *Plant Transposable Elements—Impact on Genome Structure and Function*; Grandbastien, M.-A., Casacuberta, J.M., Eds.; Springer: Berlin/Heidelberg, Germany, 2012; Volume 24, pp. 273–296.
127. McClintock, B. The significance of responses of the genome to challenge. *Science* **1984**, *226*, 792–801. [\[CrossRef\]](#)
128. Fouché, S.; Badet, T.; Oggenfuss, U.; Plissonneau, C.; Francisco, C.S.; Croll, D. Stress-Driven Transposable Element De-repression Dynamics and Virulence Evolution in a Fungal Pathogen. *Mol. Biol. Evol.* **2020**, *37*, 221–239. [\[CrossRef\]](#)
129. Shapiro, J.A. A 21st century view of evolution: Genome system architecture, repetitive DNA, and natural genetic engineering. *Gene* **2005**, *345*, 91–100. [\[CrossRef\]](#)
130. Lisch, D. How important are transposons for plant evolution? *Nat. Rev. Genet.* **2013**, *14*, 49–61. [\[CrossRef\]](#)
131. Rusterucci, C.; Espunya, M.C.; Diaz, M.; Chabannes, M.; Martinez, M.C. S-Nitrosogluthathione Reductase Affords Protection against Pathogens in Arabidopsis, Both Locally and Systemically. *Plant Physiol.* **2007**, *143*, 1282–1292. [\[CrossRef\]](#)
132. Kroumanová, K.; Kocourková, D.; Daněk, M.; Lamparová, L.; Pospíchalová, R.; Malínská, K.; Krčková, Z.; Burketová, L.; Valentová, O.; Martinec, J.; et al. Characterisation of arabidopsis flotillins in response to stresses. *Biol. Plant.* **2019**, *63*, 144–152. [\[CrossRef\]](#)
133. Daněk, M.; Valentová, O.; Martinec, J. Flotillins, Erlins, and HIRs: From Animal Base Camp to Plant New Horizons. *Crit. Rev. Plant Sci.* **2016**, *35*, 191–214. [\[CrossRef\]](#)
134. Heitz, T.; Widemann, E.; Lugan, R.; Miesch, L.; Ullmann, P.; Désaubry, L.; Holder, E.; Grausem, B.; Kandel, S.; Miesch, M.; et al. Cytochromes P450 CYP94C1 and CYP94B3 catalyze two successive oxidation steps of plant hormone jasmonoyl-isoleucine for catabolic turnover. *J. Biol. Chem.* **2012**, *287*, 6296–6306. [\[CrossRef\]](#)
135. Aubert, Y.; Widemann, E.; Miesch, L.; Pinot, F.; Heitz, T. CYP94-mediated jasmonoyl-isoleucine hormone oxidation shapes jasmonate profiles and attenuates defence responses to Botrytis cinerea infection. *J. Exp. Bot.* **2015**, *66*, 3879–3892. [\[CrossRef\]](#)
136. Kandel, S.; Sauveplane, V.; Compagnon, V.; Franke, R.; Millet, Y.; Schreiber, L.; Werck-Reichhart, D.; Pinot, F. Characterization of a methyl jasmonate and wounding-responsive cytochrome P450 of Arabidopsis thaliana catalyzing dicarboxylic fatty acid formation in vitro. *FEBS J.* **2007**, *274*, 5116–5127. [\[CrossRef\]](#)
137. Wünsche, H.; Baldwin, I.T.; Wu, J. S-Nitrosogluthathione reductase (GSNOR) mediates the biosynthesis of jasmonic acid and ethylene induced by feeding of the insect herbivore Manduca sexta and is important for jasmonate-elicited responses in Nicotiana attenuata. *J. Exp. Bot.* **2011**, *62*, 4605–4616. [\[CrossRef\]](#)

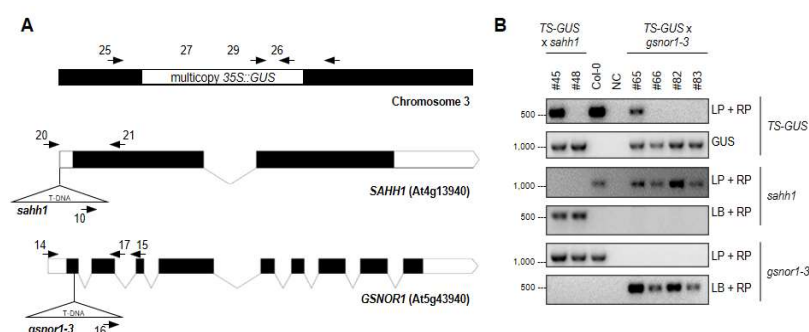


Supplementary Figure S1. Loss of AtGSNOR1 function results in an increased RSNO content under basal conditions.

(A) Phenotype of 4-week-old *Arabidopsis* mutants. **(B)** GSNOR activity was determined by NADPH consumption in the presence of GSNO ($n = 3-5$). **(C)** RSNO content was determined by triiodide-dependent ozone-based chemiluminescence ($n \geq 5$). Both analyses were measured in 4-week-old rosette leaves grown under long-day condition harvested 5 h after day-time start. Statistics: Values represent the mean $\pm$ SD of at least three independent experiments. Grubb's outlier-test ($\alpha = 0.05$) was performed. **($p < 0.01$) and ***($p < 0.001$) represents significant differences between wild-type and mutants (ANOVA with Dunnett's multiple comparisons test).

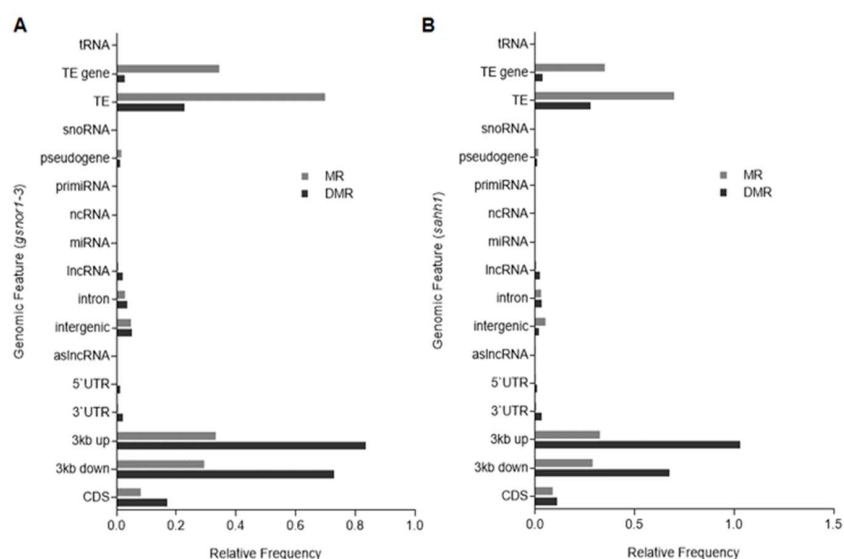


Supplementary Figure S2. SAHH1 is S-nitrosated and inhibited by GSNO. (A) *In vitro* S-nitrosation of SAHH1. 10 μ g of purified recombinant His-tagged SAHH1 (56 kDa) was treated with water, GSH, and increasing concentrations of GSNO and subjected to the biotin switch assay using ascorbate (Asc) as reducing agent. Electrophoretically separated proteins were transferred onto a nitrocellulose membrane. Ponceau S staining of blotted proteins demonstrate sample loading. Biotinylated SAHH1 was detected by an alkaline phosphatase-conjugated anti-Biotin IgG followed by BCIP/NBT chromogenic visualization. (B) Concentration dependent inhibition of SAHH1 by GSNO *in vitro*. SAHH1 produced in *E. coli* was treated with water (control), GSNO, and NEM with indicated concentration. After desalting, SAHH1 activity was measured. For restoring of SAHH1 activity, 10 mM DTT was added to the GSNO inhibited enzyme and then desalted. Statistics: Values are expressed as percentage of the control activity (at 0 mM: 0.44-0.89 nmol SAH min⁻¹ μ g⁻¹ varied among independent purification) and represent the mean $\pm$ SD of at least three independent preparations of recombinant SAHH1 (n = 3-7). Grubb's outlier-test (α = 0.05) was performed. ***(p <0.001) represents significant differences between non-treated and treated SAHH1 enzyme (ANOVA, Dunnett's multiple comparisons test). Statistical analysis was performed with GraphPad Prism version 7.05. (C) S-Nitrosation of SAHH1 (54 kDa) in wt and *gsnor1-3* leaf extracts. Leaf extracts from Col-0/wt plants (+/-GSNO) and *gsnor1-3* were subjected to the RSNO-RAC. *In vivo* S-nitrosation of SAHH1 was not detected. Total SAHH1 protein ensures equal protein loading. The assay was repeated thrice with similar results.



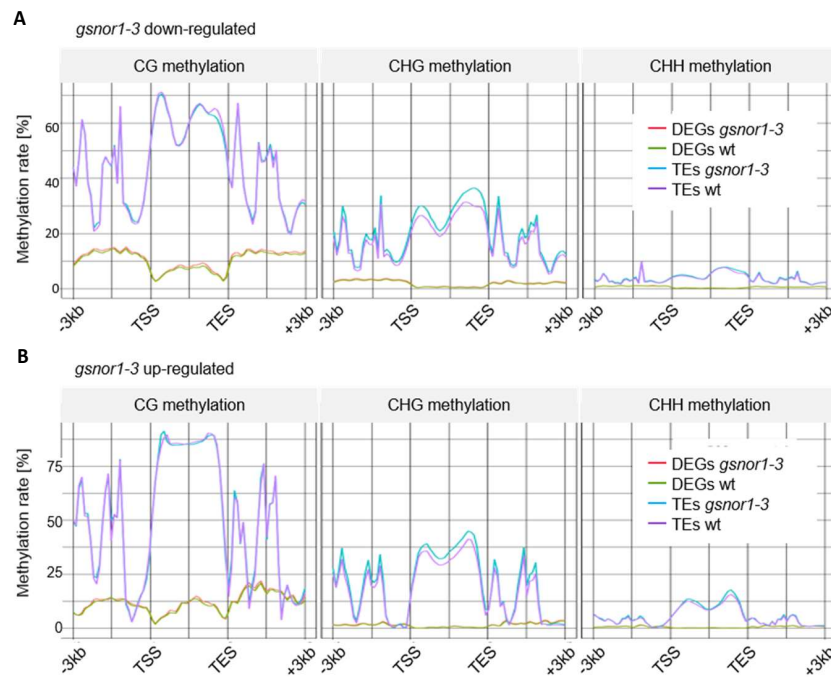
Supplementary Figure S3. PCR-based genotyping of transgenic lines harboring *TS-GUS* insertion and *sahh1* or *gsnor1-3* mutation.

(A) The diagram illustrates the positions of the single insert of a multicopy 35S::GUS transgene. Black box, Chromosome 3; white box, GUS locus. Below the position of the T-DNA insertion sites for *sahh1* and *gsnor1-3* are shown. Black box, coding region; open box, untranslated region; solid black line, intron. The T-DNA insertion of the mutant line *sahh1* are located in the 5'UTR. The position of the T-DNA insertion in *gsnor1-3* is located in the first exon. **(B)** PCR-based genotyping resulted in homogeneity for *TS-GUS x sahh1* and *TS-GUS x gsnor1-3* for #48, and #66, #82, #83, respectively. Amplicon length, primers, and PCR conditions are listed Supplemental Table S3. DNA ladder (bps) is indicated. Abbreviation: *TS-GUS*, multicopy 35S::GUS insertion; LP/RP, left/right border genomic primer; LB T-DNA border primer.



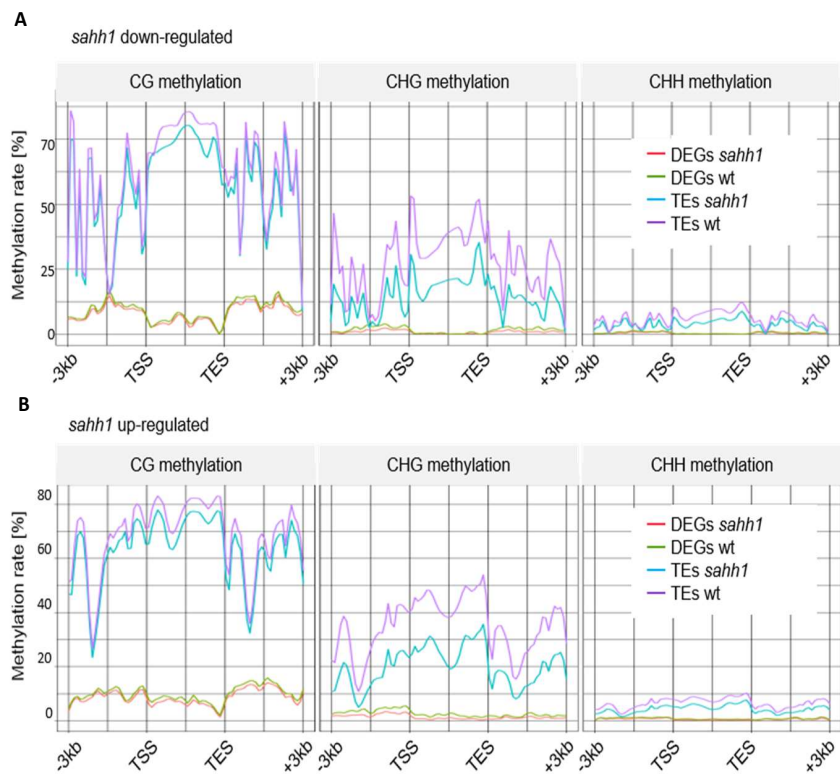
Supplementary Figure S4. Annotation of DMRs to genomic features.

Genomic context of MRs and DMRs identified in *gsnor1-3* (A) and *sah1* (B). TAIR 10 was used for annotation of genomic elements. MRs and DMRs were assigned to the following annotated elements: coding sequence (CDS), 3kb up- and 3kb down-stream of protein coding genes, 5' UTR, 3' UTR, aslncRNA, intergenic, intron, lncRNA, miRNA, ncRNA, pri-miRNA, pseudogene, snoRNA, transposable element (TE), TE gene, and tRNA.



Supplementary Figure S5. DNA methylation is poorly correlated with gene expression differences in *gsnor1-3*.

Average levels of CG, CHG, and CHH methylation over DEGs (identified as (A) down- or (B) up-regulated in *gsnor1-3*) and differentially expressed TE families as identified by RNA-seq are plotted from 3kb up- to 3kb down-stream. Abbreviations: TSS, transcriptional start site; TES, transcriptional end site; DEGs, differentially expressed genes.



Supplementary Figure S6. DNA methylation is poorly correlated with gene expression differences in *sahh1*.

Average levels of CG, CHG, and CHH methylation over DEGs (identified as **(A)** down- or **(B)** up-regulated in *sahh1*) and differentially expressed TE families as identified by RNA-seq are plotted from 3kb up- to 3kb down-stream. Abbreviations: TSS, transcriptional start site; TES, transcriptional end site; DEGs, differentially expressed genes.

Supplemental Methods

Heterologous production of AtSAHH1

AtSAHH1 was produced as N-terminal His₆-tagged fusion protein in *E. coli* BL21 (DE3) carrying the pDEST17/T7::His₆-AtSAHH1 plasmid, which encodes the full-length cDNA of AtSAHH1 (AT4G13940). An overnight pre-culture was 100-fold diluted in 500 mL of LB-Lennox media (1% tryptone (w/v), 0.5% (w/v) yeast extract, 0.5% (w/v) NaCl adjusted to pH 7 with NaOH) supplemented with 100 µg mL⁻¹ ampicillin in a 2-L Erlenmeyer baffled flask. The culture was incubated to an OD₆₀₀ of 0.6 (37°C, 180 rpm shaking rate), induced with 1 mM isopropyl β-D-1-thiogalactopyranoside (IPTG), and further cultivated at 16°C overnight. Cells were harvest by centrifugation (4,000g, 4°C, 15 min), flash frozen, and stored at -80°C until purification. For lysis, the bacteria pellet was resuspended in ice-cold lysis buffer (50 mM Tris-HCl pH 8.0, 300 mM NaCl, 20 mM imidazole, 0.5% (v/v) Triton® X-100, 1 mM β-mercaptoethanol; about 2.5 mL x g cell weight) and cells were lysed by three freeze-thaw cycles (-80°C freezer and 30°C water bath). Then, the lysate was supplemented with 1 mg mL⁻¹ lysozyme and incubated on ice with gentle shaking for 30 min. After sonification, cellular debris were removed by centrifugation (18,000g, 20 min, 4°C) and the soluble fraction was subjected to affinity chromatography using Ni-NTA agarose (Qiagen) and Econo-Pac® columns (Bio-Rad). The lysate was applied onto the column, equilibrated with lysis buffer, and washed with 10 CV of buffer A (50 mM Tris-HCl pH 8.0, 300 mM NaCl, 20 mM imidazole, 1 mM β-mercaptoethanol), 10 CV with buffer A supplemented with 1 M NaCl, and again with 10 CV of buffer A. Adsorbed proteins were eluted with buffer B (50 mM Tris-HCl pH 8.0, 300 mM NaCl, 300 mM imidazole, 20% (v/v) glycerol, 1 mM β-mercaptoethanol). The eluates were frozen in liquid nitrogen and stored at -80°C. Based on SDS-PAGE gels, fractions containing recombinant AtSAHH1 protein were pooled together and rebuffed using Zeba™ Spin columns (Thermo Scientific) equilibrated with 50 mM Tris-HCl pH 8.0 before use. The protein content was determined by Bradford [1] using the Bio-Rad Protein Assay according to the manufacturer's microplate protocol with BSA as standard.

SAHH activity assay

The activity of SAHH was determined in the SAH hydrolytic direction in the presence of excess adenosine deaminase (Ado). SAH is hydrolyzed to Hcys and adenine, which is subsequently converted into inosine and ammonia by Ado. This deamination is associated with the decrease in absorbance at 265 nm [2]. The reaction mixture contained 50 mM potassium-phosphate buffer pH 7.2, 1 mM EDTA, 0.75 U adenosine deaminase, and 2 µg purified recombinant protein in a final volume of 500 µl. After the addition of SAH to a final concentration of 100 µM, the decrease of absorbance at 265 nm at 10 s intervals for 300 s at RT in a *Ultrospec™ 3100 pro* UV/Visible spectrophotometer (Amersham Biosciences) using quartz suprasil® Hellma® Precision cells cuvettes (Hellma Analytics) was monitored. The reference sample contained water instead of SAH. The product concentrations were calculated from the slope of ΔA and $\epsilon = 8.1 \text{ mM}^{-1} \text{ cm}^{-1}$ (Ref [3]). The effect of GSNO or N-ethylmaleimide (NEM) on SAHH activity was analyzed by incubation of recombinant AtSAHH1 (in 50 mM Tris-HCl pH 8) with these compounds for 30 min at RT in the dark. Thereafter, samples were desalted using Zeba™ Spin columns equilibrated with 50 mM Tris-HCl pH 8. To assess reversibility of GSNO-dependent inhibition of recombinant AtSAHH1, 10 mM of the reducing agent dithiothreitol (DTT) was added 30 min following addition of GSNO, after which samples were incubated for 10 min prior to buffer exchange.

GSNOR activity assay

GSNOR activity in 4-week-old rosette leaves were determined as previously described [4]. Briefly, 0.5 g grinded plant material was extracted with 1 mL of extraction buffer (100 mM Tris-HCl pH 7.5, 0.1 mM EDTA, 0.2% (v/v) Triton® X-100, 10% (v/v) glycerol) on ice for 10 min with intermittent vortexing. The extracts were clarified by two centrifugation steps at 14,000g for 15 min at 4°C, and then, the protein concentration was determined according to Bradford [1] with BSA as standard. The reaction mixture

contained 20 mM Tris-HCl pH 8.0, 0.5 mM EDTA, 0.2 mM NADH, and 100 µg protein extract in a final volume of 1 mL. After the addition of GSNO to a final concentration of 500 µM, the oxidation of NADH was monitored by the decrease of absorbance at 340 nm ($\epsilon=6.22 \text{ mM}^{-1} \text{ cm}^{-1}$) at 10 s intervals for 300 s at RT in a *Ultrospec™ 3100 pro* UV/Visible spectrophotometer (Amersham Biosciences) using quartz suprasil® Hellma® Precision cells cuvettes (Hellma Analytics). The reference sample contained water instead of GSNO. The activity was expressed as consumption of nmol NADH min⁻¹ mg⁻¹.

Sodium dodecyl sulfate polyacrylamide gel electrophoresis

Sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) was performed according to Laemmli [5] as described by Sambrook and Russell [6]. Protein samples were diluted in 2x SDS sample buffer (4% (w/v) SDS, 20% (v/v) glycerol, 10% (v/v) 2-mercaptoethanol, 0.004% (w/v) bromophenol blue and 0.125 M Tris-HCl, pH 6.8.) and heated before electrophoresis for 5 min to 95°C. Proteins were separated by 10%, 12%, or 13.5% resolving gels and 5% SDS-PAGE stacking gels in SDS running buffer (25 mM Tris base, 192 mM glycine, 0.1% (w/v) SDS) on a Mini-PROTEAN® Tetra cell system (Bio-Rad). Running conditions were set to 200 V and 30 min. Proteins were visualized either by staining with *Coomassie®* Brilliant Blue G 250 or subjected to immunoblotting.

Coomassie® staining of SDS-PAGE gels

Gels were washed three times in deionized water for 2 min and then stained with Coomassie Brilliant Blue G 250 solution (0.1% Coomassie® G 250, 10% (v/v) glacial acetic acid, 45% (v/v) methanol) for 20 min while shaking. Destaining was allowed to proceed for several hours in deionized water or in destaining solution (10% (v/v) glacial acetic acid, 45% (v/v) methanol).

Immunoblotting

Electrophoretic separated proteins were blotted onto a Amersham™ Protran® 0.45 µm pore nitrocellulose membrane by semi-dry blotting using PerfectBlue™ Semi-Dry Electrobloetter Sedec™ (VWR Peqlab) and Towbin buffer [7] (192 mM glycine, 25 mM Tris, 0.1% (w/v) SDS, 20% (v/v) methanol) at 2.5 mA cm⁻² for 45 min at RT. Then blots were stained with Ponceau S to assess protein loading. After blocking in Tris-buffered saline buffer (TBS; 50 mM Tris-HCl pH 7.5, 150 mM NaCl) with 0.05% (v/v) Tween® (TBST) and 5% (v/v) BSA overnight, blots were incubated with primary antibody diluted in blocking buffer for 2 h at RT. Blots were washed thrice for 5 min in TBST. Then, the secondary antibodies in blocking buffer was added and the blot was further incubated for 1 h at RT. Thereafter, blots were washed thrice for 5 min in TBST and once with TBS. The bound secondary antibodies were detected by Western Lightning® Plus-ECL (PerkinElmer) chemiluminescence substrate and the luminograms were visualized using the Fusion FX7 imaging system (Vilber Lourmat). Signal intensities were measured using ImageJ software (National Institutes of Health) and *normalized to Ponceau S staining*. The antibodies and dilutions used were: mouse monoclonal to H3K9me2 (1:1,000; ab1220 from Abcam), mouse monoclonal to biotin (1:5,000; A6561 Sigma), anti-rabbit IgG (1:2,500; W4011, Promega) and anti-mouse IgG (1:2,500; W4021, Promega), rabbit polyclonal to AtSAHH1 (1:10,000; Ref. [8]; kindly provided by Barbara Moffat).

Detection of S-nitrosated proteins

The biotin switch technique (BST; Ref. [9]) or the resin-assisted capture of S-nitrosothiols assay (RSNO-RAC; Ref. [10]) were used to analyze S-nitrosation of proteins. 10 µg of recombinant AtSAHH1 in HEN buffer (100 mM Hepes-NaOH pH 7.7, 1 mM EDTA, 0.1 mM neocuproine) was treated with 50 µM, 250 µM, 500 µM GSNO or 500 µM GSH in a final volume of 100 µL for 30 min at RT in the dark with intermittent inverting of the tubes. Free thiols were blocked with 20 mM S-methylmethanethio-sulfonate (MMTS) in the presence of 2.5% (w/v) SDS at 300 rpm for 20 min at 50°C on a thermoshaker in the dark. After acetone precipitation to remove excess of GSNO and MMTS, the protein pellets were washed thrice with ice-cold acetone (10,000g for 10 min at 4°C), air-dried, and then resuspended in 32.5 µL of HENS buffer (HEN supplemented with 1% (w/v) SDS). The labeling reaction was started by the reduction of S-nitrosothiol

groups with sodium ascorbate (1 mM final concentration) to generate thiols and O-nitrosoascorbates via a transnitrosation reactions. The nascent reduced thiols, which were originally S-nitrosated, were then biotinylated by adding 1 mM biotin-HPDP. The labeling reaction (50 μ L final volume) was allowed to proceed at 300 rpm for 1 h at RT using a thermoshaker in the dark. After acetone precipitation, the biotin-labeled proteins were resuspended in 45 μ L non-reducing sample buffer (0.06 M Tris-HCl, pH 6.8, 10% (v/v) glycerol, 0.008% (w/v) bromophenol blue) and separated on a 10% SDS-PAGE gel without boiling prior to loading. Biotinylated proteins were detected with an anti-biotin antibody conjugated with alkaline phosphatase and 5-Bromo-4-chloro-3-indolyl phosphate/Nitro blue tetrazolium (BCIP/NBT) chromogenic visualization. *In planta*, S-nitrosation of AtSAHH1 was investigated in 4-week-old *Arabidopsis* rosette leaves homogenized in HEN-T buffer (HEN supplemented with 0.2% (v/v) Triton® X-100 and protease inhibitor cocktail) in a ratio of 1:2. The extract was centrifuged twice at 18,000g for 20 min at 4°C. After desalting using Zeba™ Spin columns equilibrated with HEN buffer, the protein content was assayed by Bradford [1] using the Bio-Rad Protein Assay according to the manufacturer's microplate protocol with BSA as standard. For *in vitro* S-nitrosation, extracts (1.5 mg; 1.13 μ g μ L⁻¹) were incubated with 0.5 mM GSNO or water as mock for 30 min at RT in the dark with frequent inverting. Next, free cysteine thiols were blocked by S-methylthiolation with 20 mM MMTS in the presence of 2.5% SDS at 300 rpm for 20 min at 50°C on a thermoshaker in the dark in a final volume of 1.5 mL. Excess of GSNO and MMTS was removed by precipitation with two volumes of ice-cold acetone for 20 min at -20°C. After centrifugation (10,000g for 10 min at 4°C), the protein pellet was washed three times with 70% (v/v) acetone, air-dried, and resuspended in HENS buffer (1 mL per mg protein). Afterwards 30 μ L of the protein suspension were taken for analysis of total protein input. Next, 16 mg thiopropyl sepharose 6B per sample (equal to 50 μ L bed volume) was equilibrated in excess HEN buffer for 15 min, washed thrice with HEN buffer by centrifugation (1,000g for 1 min) and resuspended by the addition of 150 μ L HEN buffer. The protein suspension was added to 200 μ L slurry (equal to 50 μ L bed volume) in the presence of 50 mM ascorbate (prepared in HEN buffer) to a final volume of 2.3 mL. Thus, nascent SNO-proteins bind after their reduction to thiols with ascorbate to the thiol-reactive resin. Ascorbate was omitted for negative controls. Protein capture was allowed to proceed for 3 h at RT in the dark using an overhead shaker. To remove non-bond proteins, the resin was washed four times with 3 mL of HENS buffer and two times with 2 mL of HENS/10 buffer (HENS diluted 1:10). Centrifugation steps to collect beads were performed at 500g for 1 min. Originally S-nitrosated proteins were eluted with one bed volume of HENS/10 supplemented with 100 mM β -mercaptoethanol for 20 min at 22°C on a thermo-shaker (1500 rpm), separated on a 13.5% SDS-PAGE, and analyzed by immunoblotting with an anti-AtSAHH1 antibody.

GUS activity staining

GUS activity was detected by histochemical staining according to Ref. [11] with modifications. *A. thaliana* line L5 harboring multiple copies of a 35Spro::GUS marker gene (*TS-GUS*; Morel et al. 2000) plantlets were immersed in fixing solution (50 mM sodium phosphate buffer pH 7.0, 0.05% (v/v) Triton® X-100, 0.5% (v/v) formaldehyde), vacuum infiltrated, and washed thrice with 50 mM sodium phosphate buffer pH 7.0. Then the plantlets were immersed into GUS staining solution (50 mM sodium phosphate buffer pH 7.0, 0.1% Triton® X-100, 1 mM potassium ferrocyanide, 1 mM potassium ferricyanide, 1 mM 5-bromo-4-chloro-3-indolyl-D-glucuronic acid) and incubated at 37°C overnight. Plantlets were washed once with 70% (v/v) ethanol and incubated in 70% (v/v) ethanol at 80°C to remove chlorophyll for 10 min.

Quantification of S-nitrosothiols

The S-nitrosothiol level in 4-week-old rosette leaves were determined by triiodide-dependent ozone-based chemiluminescence using the Nitric Oxide Analyzer Sievers® 280i from GE Healthcare. This method is based on the reduction of nitrite and RSNO by triiodide to NO, which reacts with ozone to form the excited state of nitrogen dioxide (NO₂^{*}) and O₂. Upon decay to its ground state, NO₂^{*} emits a photon which is detected by a photomultiplier [13]. The intensity of emitted light is directly proportional to the amount of NO. In short, rosette leave extracts were prepared by homogenization of 500 mg ground plant material in

two volumes of 1x phosphate buffered saline (PBS; 137 mM NaCl, 2.7 mM KCl, 4.3 mM Na₂HPO₄, 1.47 mM KH₂PO₄, pH 7.4) and incubation on ice for 10 min with intermittent vortexing. After centrifugation at 18,000g for 15 min at 4°C, the supernatant was used for further analysis. For RSNO measurements, extracts were pre-treated with 5% sulfanilamide (w/v, in 1 M HCl) at a ratio of 9:1 (extract : sulfanilamide) to scavenge nitrite and 200 µL were injected into the reaction vessel containing acidic triiodide as reducing agent (28.5 mM I₂, 66.9 mM KI, 77.7% glacial acetic acid; 30°C). The peak area integration and quantification of RSNO content were performed with Sievers® NOA Analysis™ software (GE Healthcare) using nitrite standards and normalized to the protein content assayed according to Bradford [1].

References

1. Bradford, M.M. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal. Biochem.* **1976**, *72*, 248–254, doi:https://doi.org/10.1016/0003-2697(76)90527-3.
2. Fujioka, M.; Takata, Y. S-Adenosylhomocysteine hydrolase from rat liver. Purification and some properties. *J. Biol. Chem.* **1981**, *256*, 1631–1635, doi:10.1139/o82-016.
3. Takata, Y.; Yamada, T.; Huang, Y.; Komoto, J.; Gomi, T.; Ogawa, H.; Fujioka, M.; Takusagawa, F. Catalytic mechanism of S-adenosylhomocysteine hydrolase. Site-directed mutagenesis of Asp-130, Lys-185, Asp-189, and Asn-190. *J. Biol. Chem.* **2002**, *277*, 22670–22676, doi:10.1074/jbc.M201116200.
4. Sakamoto, A.; Ueda, M.; Morikawa, H. Arabidopsis glutathione-dependent formaldehyde dehydrogenase is an S-nitrosogluthathione reductase. *FEBS Lett.* **2002**, *515*, 20–24, doi:10.1016/S0014-5793(02)02414-6.
5. Laemmli, U.K. Cleavage of Structural Proteins during the Assembly of the Head of Bacteriophage T4. *Nature* **1970**, *227*, 680–685, doi:10.1046/j.1365-313x.1998.00343.x.
6. Sambrook, J.; Russell, D.W. *Molecular cloning: a laboratory manual*; 3rd ed.; Cold Spring Harbor Laboratory Press: New York, 2001; ISBN 0879695773 (paperback).
7. Towbin, H.; Staehelin, T.; Gordon, J. Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: procedure and some applications. *Proc. Natl. Acad. Sci. U. S. A.* **1979**, *76*, 4350–4, doi:10.1002/bies.950190612.
8. Pereira, L.A.; Todorova, M.; Cai, X.; Makaroff, C.A.; Emery, R.J.; Moffatt, B.A. Methyl recycling activities are co-ordinately regulated during plant development. *J. Exp. Bot.* **2007**, *58*, 1083–1098, doi:10.1093/jxb/erl275.
9. Jaffrey, S.R.; Snyder, S.H. The Biotin Switch Method for the Detection of S-Nitrosylated Proteins. *Sci. Signal.* **2001**, *2001*, p11–p11, doi:10.1126/stke.2001.86.p11.
10. Forrester, M.T.; Thompson, J.W.; Foster, M.W.; Nogueira, L.; Moseley, M.A.; Stampler, J.S. Proteomic analysis of S-nitrosylation and denitrosylation by resin-assisted capture. *Nat. Biotechnol.* **2009**, *27*, 557–559, doi:10.1038/nbt.1545.
11. Jefferson, R.A.; Kavanagh, T.A.; Bevan, M.W. GUS fusions: beta-glucuronidase as a sensitive and versatile gene fusion marker in higher plants. *EMBO J.* **1987**, *6*, 3901–7, doi:10.1073/pnas.1411926112.
12. Morel, J.-B.; Mourrain, P.; Béclin, C.; Vaucheret, H. DNA methylation and chromatin structure affect transcriptional and post-transcriptional transgene silencing in Arabidopsis. *Curr. Biol.* **2000**, *10*, 1591–1594, doi:http://dx.doi.org/10.1016/S0960-9822(00)00862-9.
13. Piknova, B.; Schechter, A.N. Measurement of Nitrite in Blood Samples Using the Ferricyanide-Based Hemoglobin Oxidation Assay. *Methods Mol. Biol.* **2011**, *704*, 39–56, doi:10.1007/978-1-61737-964-2_4.

5 ARADEEPOPSIS, an Automated Workflow for Top-View Plant Phenomics using Semantic Segmentation of Leaf States

status	<i>published</i> The Plant Cell 32, 12, 3674–3688 (2020) doi:10.1105/tpc.20.00318
authors	Patrick Hüther* (P.H.), Niklas Schandry* (N.S.), Katharina Jandrasits (K.J.), Ilja Bezrukov (I.B.) and Claude Becker (C.B.) <i>*equal contributions</i>
contributions	P.H., N.S., and C.B. designed the research and wrote the article; P.H., N.S., and K.J. performed the research; P.H. wrote the pipeline; N.S. and K.J. performed the phenotyping experiment and generated the imagedata; I.B. provided additional image and segmentation data; P.H. and N.S. analysed the data.

The Plant Cell, Vol. 32: 3674–3688, December 2020, www.plantcell.org © 2020 The authors.



LARGE-SCALE BIOLOGY ARTICLE

ARADEEPOPSIS, an Automated Workflow for Top-View Plant Phenomics using Semantic Segmentation of Leaf States^[CC-BY]

Patrick Hüther,^{a,1,2} Niklas Schandry,^{a,b,1} Katharina Jandrasits,^a Ilja Bezrukov,^c and Claude Becker^{a,b,2}

^aGregor Mendel Institute of Molecular Plant Biology (GMI), Austrian Academy of Sciences, Vienna BioCenter (VBC), 1030 Vienna, Austria

^bGenetics, Faculty of Biology, Ludwig-Maximilians-University München, 82152 Martinsried, Germany

^cDepartment of Molecular Biology, Max Planck Institute of Developmental Biology, 72076 Tübingen, Germany

Linking plant phenotype to genotype is a common goal to both plant breeders and geneticists. However, collecting phenotypic data for large numbers of plants remain a bottleneck. Plant phenotyping is mostly image based and therefore requires rapid and robust extraction of phenotypic measurements from image data. However, because segmentation tools usually rely on color information, they are sensitive to background or plant color deviations. We have developed a versatile, fully open-source pipeline to extract phenotypic measurements from plant images in an unsupervised manner. ARADEEPOPSIS (<https://github.com/Gregor-Mendel-Institute/aradeepopsis>) uses semantic segmentation of top-view images to classify leaf tissue into three categories: healthy, anthocyanin rich, and senescent. This makes it particularly powerful at quantitative phenotyping of different developmental stages, mutants with aberrant leaf color and/or phenotype, and plants growing in stressful conditions. On a panel of 210 natural *Arabidopsis* (*Arabidopsis thaliana*) accessions, we were able to not only accurately segment images of phenotypically diverse genotypes but also to identify known loci related to anthocyanin production and early necrosis in genome-wide association analyses. Our pipeline accurately processed images of diverse origin, quality, and background composition, and of a distantly related Brassicaceae. ARADEEPOPSIS is deployable on most operating systems and high-performance computing environments and can be used independently of bioinformatics expertise and resources.

Introduction

The Phenotyping Bottleneck in Plant -Omics Studies

Over the last decades, molecular techniques have steadily increased in throughput while decreasing in cost. A prime example is nucleic acid sequencing, which has followed a trend analogous to Moore's law in computer science. However, phenotyping methods aimed at determining the physical shape of an organism and measuring morphological parameters have not kept up with this pace, which has caused a "phenotyping bottleneck" (Furbank and Tester, 2011) in the design and execution of scientific studies. This phenotyping bottleneck constitutes a major challenge in plant biology as well, due to two major issues: image acquisition and data analysis. Data acquisition in most plant phenotyping scenarios involves acquiring standardized images of plants growing in a controlled environment. Plant phenotyping requires space and

a dedicated infrastructure that can accommodate the developmental and growth transitions that occur over the life of a plant. Moreover, plant development needs to be phenotyped over relatively long periods of time. For example, in the case of the model plant *Arabidopsis* (*Arabidopsis thaliana*), a relatively small and fast-growing species, a phenotyping experiment typically runs for several weeks or months, depending on the phenotype of interest and growth conditions. In their vegetative phase, that is, before they produce shoots, flowers, and seeds, *Arabidopsis* plants grow as relatively small, flat rosettes and can therefore be considered two-dimensional. Because the entire plant is visible from the top during this phase, top-view phenotyping is a straightforward method to determine size, growth rate, and leaf development. While this type of high-throughput image acquisition has become almost trivial, robustly and faithfully extracting meaningful data from these images has not.

Indeed, overcoming the data acquisition challenge, by means of a dedicated plant phenotyping facility for example, does not mitigate the second cause of the phenotyping bottleneck: data analysis. In the context of image-based phenotyping data, analysis includes automated image processing for object detection, object segmentation, and extraction of quantitative phenotypic traits.

Challenges in Image Analysis

The first key step from images to quantitative phenotypic data is also the most difficult: defining the area in the image that depicts

¹ Patrick Hüther and Niklas Schandry contributed equally to this work.

² Address correspondence to patrick.huether@gmi.oew.ac.at or claudbecker@gmi.oew.ac.at.

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantcell.org) is: Claude Becker (claudbecker@gmi.oew.ac.at).

^[CC-BY]Article free via Creative Commons CC-BY 4.0 license. www.plantcell.org/cgi/doi/10.1105/tpc.20.00318



IN A NUTSHELL

Background: To understand the function of a gene, we need to be able to link genetic information and phenotype. For many research questions in plant biology, it is therefore important to accurately measure the size, shape, and color of the plant. Often, plant phenotyping means collecting vast amounts of images, which one then needs to process and annotate to extract plant morphometric information. In recent years, the speed and dimensions at which we can decode genomic information have increased exponentially; by contrast, the bottleneck of large-scale plant research now often lies in reliable automated image processing. Our goal was to develop a pipeline that accurately annotates plants in top-view images and extracts phenotypic parameters while requiring minimal user input.

Question: We set out to build an unsupervised machine learning-based pipeline for plant phenotyping that would not only segment plants from top-view images but would also return measurements for a range of phenotypic traits. The pipeline, named ARADEEPOPSIS, had to be able to process large datasets, be accessible to non-expert users, and run on common computer environments.

Findings: We show that one can use a relatively small number of manually-annotated images of *Arabidopsis thaliana* to successfully re-train an already established Deep Neural Network, underscoring the power of transfer-learning approaches to address challenges in plant phenotyping. We demonstrate the versatility of the pipeline by analyzing a collection 150,000 images across plant development, not only correctly segmenting rosette areas but also automatically classifying leaves based on their health status. Our pipeline thus allows automated segmentation of anthocyanin-rich and senescent areas, which we were then able to use in genome-wide association studies to link these phenotypes to standing genetic variation in the species.

Next steps: ARADEEPOPSIS was designed to segment plant rosettes specifically from top-view images of individual plants. The pipeline illustrates that transfer learning is a powerful approach to solve plant phenotyping problems. In the future, we wish to expand image segmentation to recognize a wider array of plant species and tissues. Moreover, we are working on routines to identify plants from images that contain more than one individual.



the object of interest, in this case a plant. On a small set of images, this segmentation can be done manually by delineating the plant object using software such as *ImageJ* (Schindelin et al., 2012; Schneider et al., 2012). On data sets consisting of hundreds or thousands of images, which are typical for experiments from phenotyping platforms, this segmentation task must be automated, both to speed up the process and to neutralize user bias. Commonly used software for segmenting plant objects from digital images rely on color information. In simple terms, such color information is stored in tables, with information on each pixel stored in a corresponding cell. While grayscale images are stored in a single table, where each cell contains the grayscale value of the corresponding pixel, color image information is stored in several such tables. For images using the very common additive Red-Green-Blue (RGB) color model, three separate tables store information for the red (R), green (G), and blue (B) color channel intensities for each pixel. A very simple approach to differentiate plants from background is to assume that, since plants are green, one can assign all pixels that pass a certain threshold in the green channel to the object 'plant' and ignore all other pixels. This approach is called binary thresholding and works well if, and only if, the assumption is correct that all plants in the experiment are green and that the remaining area of the image is not. In reality, this assumption is often violated: plants can produce high quantities of anthocyanins under certain circumstances, which give plants a red or purple hue. Plants may also develop chlorosis or become senescent, in which case they turn yellow or brown and, ultimately, white. Moreover, as they grow larger over the course of an experiment, the leaves of neighboring plants can protrude into the image area of the monitored plant, thus creating additional green areas that should not be taken into account. Even the background color can fluctuate, either because of constraints in the

experimental setup, or because other organisms, such as algae, start growing on the soil surface. In summary, color-based segmentation is sensitive to image parameters and developmental stage, which are difficult to control in a biological experiment, and therefore often needs to be verified by laborious and time-consuming visual inspection of the individual images to identify and correct erroneous segmentation and artifacts. We therefore sought to develop an alternative, unsupervised, high-throughput approach that is robust against variation in color parameters and image characteristics.

Machine Learning Methods in Plant Phenotyping

Alternative approaches to object segmentation that solve the aforementioned problems of color-based segmentation are available: they are based on machine learning methods and include Gaussian mixture models (Rother et al., 2004) or Naïve Bayes classifiers (Gehan et al., 2017). While these approaches are highly flexible, their correct implementation on new data sets requires substantial programming knowledge and deeper understanding of the underlying statistical principles.

More recently, successful alternative approaches to segment particular classes of objects from images have come from the deep learning field. Convolutional neural networks have proven invaluable for image classification and segmentation tasks (Lecun et al., 1998; Krizhevsky et al., 2012; He et al., 2016; Chollet, 2017). They perform well on data that has a local structure, which is inherently the case in images where values of neighboring pixels tend to be highly correlated and contain recurring structures such as corners and edges. These models are supervised, meaning that they are provided with a set of manually annotated images containing the "ground truth" for each pixel, from which the model will

attempt to derive rules for the classification of pixels based on the training data set. This training process is iterative and usually performed over many thousands of iterations, during which the model attempts to map input images to their corresponding ground truth annotations. In brief, during training, a loss function is calculated to estimate the error between input and output, and the model subsequently tries to minimize this error via back-propagation (Rumelhart et al., 1986). Weight matrices throughout the network layers are updated along a gradient, following the chain rule to calculate partial derivatives of the loss function with regard to the layer weights. Due to the nonconvex nature of the loss function, error minimization typically reaches only local minima, requiring careful selection of model parameters and a large and diverse set of training data to avoid overfitting. The latter is usually scarce because ground truth data for supervised learning have to be generated by meticulous manual annotation. Depending on the nature of the desired feature to be extracted, such a task is typically labor intensive and time consuming due to the large number of data points required to train a well-generalizing deep learning model de novo, and hence exceeds the capacity of many experimental research labs.

Such models have been successfully applied to plant phenotyping for both aerial and root tissue (Pound et al., 2017; Ubbens et al., 2018; Yasrab et al., 2019; Wang et al., 2020). The Deep Plant Phenomics (DPP) platform provides a framework to either build models suitable for phenotyping tasks or use pretrained models (Ubbens and Stavness, 2017). Because these models are trained from scratch, their ability to generalize to unseen data is limited by the amount of annotations contained in the training set, even if data augmentation such as random cropping, contrast, or brightness adjustments are performed. Despite the lowered bar of entry provided by DPP, training of specialized models still requires programming knowledge, which may prevent nonexpert users from applying these tools to their data.

Here, we use an alternative approach toward the application of convolutional neural networks in plant phenomics. Our aim was to lower the amount of required training data, while at the same time reaching high accuracy across images sourced from different imaging platforms. To this end, we make use of a well-established large network architecture that had already been trained on the ImageNet (Deng et al., 2009) data set, consisting of millions of annotated images. Our approach, named ARADEEPOPSIS (ARAbidopsis DEEP-learning-based OPTimal Semantic Image Segmentation), is an image analysis pipeline for unsupervised plant segmentation and phenotypic trait extraction from large image data sets. Rather than developing a new model for plant segmentation, we used transfer learning by centering our pipeline around the well-established deep-learning model DeepLabV3+, which we re-trained for segmentation of Arabidopsis rosettes in top view. Besides accurately identifying image areas containing green plants (rosettes), our pipeline also faithfully phenotypes very late developmental stages, in which plants often undergo color changes. ARADEEPOPSIS can discriminate three different health states of leaves (green/healthy, high in anthocyanins, and senescent) during image segmentation, during which it is able to accurately segment and quantify plants showing altered phenotypes arising from developmental transitions or physiological responses to environmental stresses. We show how

ARADEEPOPSIS can be applied to reliably segment Arabidopsis rosettes independent of shape, age, health state, and background, which is not possible with color-based segmentation algorithms. By noninvasively extracting color index data from the segmented leaf area, our tool delivered highly resolved quantitative data that we successfully applied to a genome-wide association (GWA) analysis of anthocyanin content. ARADEEPOPSIS succeeds in delivering quantitative data on the notoriously difficult phenotype of plant senescence from the analysis of plant images, from which we identified genetic variants that drive premature senescence. Because Arabidopsis is a widely used model species in molecular plant biology and genetics, we centered our pipeline on this species. Additionally, we show that it can accurately segment other species.

ARADEEPOPSIS requires minimal user input compared to other tools that offer semi-automated analysis; it is open source and available for free. ARADEEPOPSIS provides a convenient all-in-one solution for machine learning-based image analysis and trait-extraction and significantly lowers the bar to nonbioinformaticians who may wish to apply machine learning for plant phenotyping. Because the pipeline is written in *Nextflow* and supports containerization, researchers with little computational background can install and run it on their personal computer, cloud infrastructure, or in a high-performance computing environment. ARADEEPOPSIS is available at <https://github.com/Gregor-Mendel-Institute/aradeepopsis>.

Results

Color-Guided Segmentation Can be Misguided by Image and Object Parameters

We wished to explore the performance and accuracy of classical color-based segmentation and of a self-learning algorithm for segmenting Arabidopsis rosettes from a large, automated phenotyping experiment. We monitored the growth of 210 Arabidopsis accessions from the 1001 genomes panel (The 1001 Genomes Consortium, 2016) in natural soil, under climate-controlled conditions, with six replicates per genotype, from seed to flowering stage. Using an automated phenotyping system, we recorded top-view images twice per day, cropped them to frames containing single pots, and subjected them to rosette area measurements. We first used *LemnaGrid* (Lemnatec), which, similar to other integrated pipelines that perform (semi)-supervised segmentation, relies on color channel information to segment plants in top-view images and is commonly used for Arabidopsis phenotyping (Arvidsson et al., 2011; Minervini et al., 2017). The implementation of this pipeline resulted in the accurate segmentation of young and healthy plants that were mainly composed of green tissue (Figure 1). However, because the plants in our data set were grown over a long period in natural soil, many accumulated high levels of anthocyanins in their leaves, either because of genetic determinants or as a general stress response to the natural soil and its abiotic or biotic composition; others showed an onset of senescence at later stages of the experiment. For such images, segmentation often failed completely or resulted in only partial or inaccurate segmentation of the plant (Figure 1),

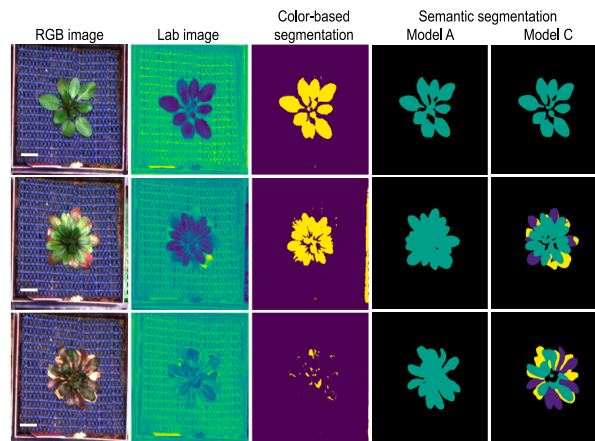


Figure 1. Performance of Color-Based Versus Semantic Segmentation.

The left-most column shows the original RGB images of a representative *Arabidopsis* plant from our phenotyping experiment at three different developmental stages. The second column shows the same images transformed into the 'Lab' color space, which enhances contrast between green and other colors. Binary thresholding (third column), based on Lab input and used for color-based segmentation, results in difficulties to correctly segment older or anthocyanin-containing plants. Semantic segmentation by our models A and C is insensitive to color and contrast of the plant (two right-most columns). For an explanation of the model C color code, see Figure 2. Scale bar = 1 cm.

showing that color-based segmentation is sensitive to deviations from the expected color composition of the object to be detected.

Training of a Proof-of-Principle Model

We hypothesized that self-learning algorithms would achieve higher accuracy while at the same time be more robust against shifts in color patterns. In a process referred to as transfer learning, trained model architectures that do well at pixel classification can be re-trained on new data that contain new classes while retaining already trained weights in layers that extract low-level features such as shapes and edges. In other words, a model may be re-trained on plant rosettes even when the initial data set the model was trained on does not contain plant rosettes. Model design is a complex task, and experts in the field of machine learning have invested considerable effort into establishing sophisticated model architectures that can be built upon. Moreover, because re-training a model mainly consists of updating the last layer of the model, transfer learning on a model that already performs well at image segmentation is much faster and requires less training data than designing and training a new model from scratch, since the established model is already able to extract features from image data and is merely learning a new classification output (Razavian et al., 2014).

To test whether transfer learning of the well-established segmentation model DeepLabV3+ would result in a model that could faithfully extract *Arabidopsis* rosettes from top-view images, independently of developmental stage or phenotype, we generated a 'ground-truth' training set. We generated a small data set of 300 images, in which we manually annotated the rosette area, deliberately excluding senescent leaves. From these 300

annotations, we randomly selected 80% (240 images) for model training and kept the remaining 20% for evaluating the accuracy of the new model.

After training, this model (referred to hereafter as model A) performed well on plants at various developmental stages and with different leaf coloring (Figure 1), which encouraged us to generate a much larger ground-truth data set for more fine-grained models (<https://doi.org/10.5281/zenodo.3946393>).

Advanced Models for Differentiated Plant Area Classification of Individual Leaves

From the example images shown in Figure 1, it is apparent that plant health changes not only in the context of the whole rosette but also between individual leaves or parts of leaves, with some accumulating high anthocyanin levels and others entering senescence. We therefore asked if we could train a model that would semantically extract such features from leaves and assign them to different classes, depending on their phenotypic appearance. Not only would such a model allow finer resolution when assessing the overall health of the plant, it would also enable segmenting differently colored areas.

Reasoning that it should be possible to segment senescent leaves separately, we generated a second, larger, and more finely annotated training data set, consisting of 1375 manually annotated top-view images of *Arabidopsis* plants from the phenotyping experiment described above, from which we again withheld 20% for later validation. The images used for annotation were selected semi-randomly such that they covered various developmental stages and included healthy-looking as well as stressed plants from different genotypes that exhibited altered color profiles due

3678 The Plant Cell

to anthocyanin accumulation or senescence, and that were diverse in leaf-morphology. Instead of manually annotating the entire rosette, as we had done for the initial training set, we annotated all leaves of each rosette individually and manually assigned each leaf to one of three different classes, depending on its appearance: green; anthocyanin-rich; senescent or dead. From these annotations, we generated image masks for two additional models complementing the initial one-class model A. For the two-class model B, we classified senescent (class_senesc) versus nonsenescent leaves (class_norm), whereas the three-class model C was trained to differentiate between senescent (class_senesc), anthocyanin-rich (class_antho), and green (class_norm) areas (Figure 2A). We then used these three annotated sets of masks for transfer learning of a publicly available *xception65* (Chollet, 2017) based DeepLabV3+ checkpoint that had been pretrained on the ImageNet data set (see Methods and <https://doi.org/10.5281/zenodo.3946618>; Chen et al., 2017; Chen et al., 2018).

After training of all three models had completed, we assessed their performance according to the mean intersection over union (mIoU) for all pixel-level classifications, which is defined

as the mean fraction of true positives divided by the sum of true positives, false positives, and false negatives over all annotated classes. When applied to the validation data set, models A, B, and C reached an mIoU of 96.8%, 84.0%, and 83.1%, respectively.

Next, we compared how each model performed at segmenting the 148,792 rosette images comprising our data set. First, we asked how much the different classes contributed to the area segmented as 'plant'. Averaged across all images, model A, which ignores senescent leaves, classified 12.5% of image area as plant tissue (Figure 2B). Models B and C both classified ~14% of image area as plant tissue (including senescent areas ignored by model A; Figure 2B). The fraction of class_senesc segmentation was almost identical in models B and C. To further test whether our most complex model, model C, performed well, we analyzed the ratio of the three classes over the course of the experiment. As expected, we saw a sharp increase in anthocyanin-rich area beginning 30 d after sowing, followed by an increase in senescent segmentations 10 d later (Figure 2C). Relative classification of green tissue decreased accordingly. This observation demonstrated that model C successfully captured an increase in

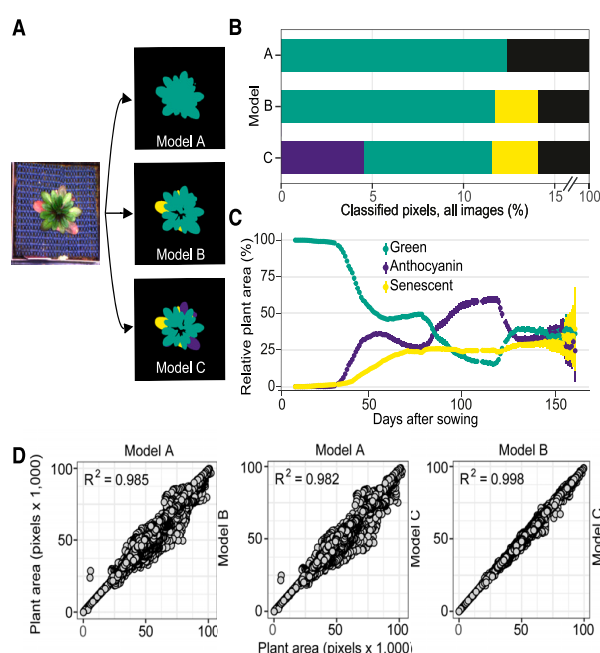


Figure 2. The Three Default Models Available Through ARADEEPOPSIS.

(A) Example of segmentation results from the one-class model A, the two-class model B, and the three-class model C on a single *Arabidopsis* individual. (B) Comparison of the percentage of classified pixels across all 148,792 images between the three available models (colors as in C; black = background). (C) Relative number of pixels classified as green, anthocyanin-rich, or senescent by model C over time. The mean percentage of pixels assigned to each class is shown per day. Error bars indicate 95% confidence intervals. (D) Scatterplots showing pairwise correlations between measurements of plant area (includes all leaf states except senescence) between models A, B, and C.

Downloaded from <https://academic.oup.com/plcell/article/32/12/3674/6118590> by guest on 14 July 2021

senescence as plants grew older, accurately reflecting plant development.

Next, we determined whether the three models behaved differently in segmenting the combined rosette area. Using the total pixel area classified as leaves from each image, we found that models B (two-class) and C (three-class) displayed the strongest correlation ($R^2 = 0.998$; Figure 2D). Model A, our one-class model trained on annotations of whole rosettes rather than individual leaves, correlated slightly less well with models B and C ($R^2 = 0.985$ and 0.982 , respectively) and had a tendency to segment larger areas than either of the other two models (Figure 2D). Visual inspection of overlays of the segmentations produced by model A with the original image revealed that segmentations produced by model A frequently included areas between leaves that were ignored by models B and C, explaining the larger rosette sizes measured by model A (see exemplary segmentations in Figure 2A). We believe that these differences are related to the different annotation strategies (whole rosettes in model A vs. more refined per-leaf annotations for the larger data sets in models B and C). When generating the ground-truth training set for models B and C, we noticed how nontrivial it was to make consistent decisions as to when a leaf should be annotated as senescent or rich in anthocyanins, as the transitions are gradual, hence classification can become subjective. This lack of total precision during annotation of the training set might also explain the decrease in mIoU scores that we observed with the two- and three-class models. We attempted to mitigate user bias by having several individuals contributing to generating the annotations for a subset of images, thus ensuring that the model would faithfully learn the features of interest.

Extraction of Morphometric and Color Index Traits from Segmented Images

Next, we used the segmentation masks obtained from the models created by transfer learning to extract morphometric and color index traits associated with each plant (Figure 3), using the Python library *scikit-image* (van der Walt et al., 2014; Del Valle et al., 2018). Besides the total number of pixels annotated as 'plant', the pipeline returns the equivalent diameter, that is, the diameter of a circle with the same number of pixels as the plant area. The major and minor axes of an object-surrounding ellipse are a measure of the aspect ratio of the plant and inform on whether the rosette is rather round or elongated along one axis (Figure 3A). The area of a convex hull surrounding the object is a measure of plant size; the solidity of the plant object is calculated by dividing the area of the plant by the area of the convex hull. Finally, dividing the plant area by the area of its bounding box, that is, a minimal box enclosing the object, is representative of the overall extent of the plant (Figure 3A). Depending on the type of downstream analyses, these indirect object parameters can be used as proxies for overall or specific growth.

The segmentation masks are also used to extract color index information from the plant area (Figure 3B). These traits are based on the intensity of each RGB color channel per pixel. Simple color index traits are, for example, the average intensity of the green channel over all pixels classified as 'plant.' We have implemented the color index traits described by Del Valle et al. (2018) in our pipeline (Figure 3C). These traits are calculated for each class individually and for a compound class termed "plant_region," which contains both class_norm and class_antho (only for model

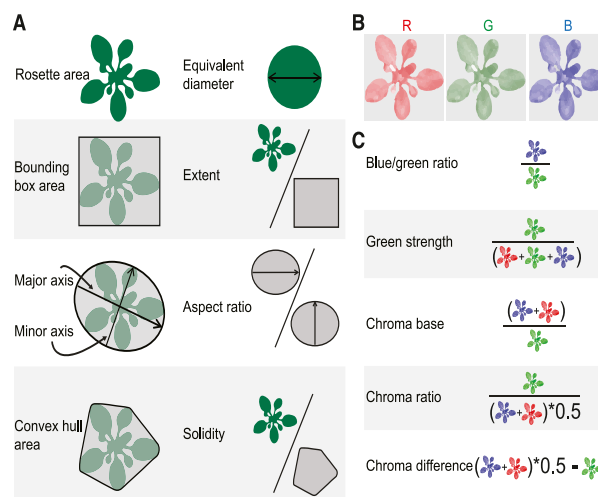


Figure 3. Morphometric and Color Index Measurements Extracted from Segmentation Masks.

(A) Morphometric traits extracted using the Python library *scikit-image* (van der Walt et al., 2014).

(B) Separation of the segmented plant image into red (R), green (G), and blue (B) color channels.

(C) Color channel indices calculated as described in Del Valle et al. (2018). The differently colored plants represent mean values of the different color channels; see (B).

3680 The Plant Cell

C). Details on the color index traits are shown in Figure 3C and provided in Methods.

The ARADEEPOPSIS Pipeline

The ARADEEPOPSIS (ARAbidopsis DEEP-learning-based OPTimal Semantic Image Segmentation) pipeline integrates the unsupervised image segmentation with the trait extraction outlined above. The segmentation results are easily browsable with a custom *Shiny* application, allowing for straightforward quality assessment of the obtained results. ARADEEPOPSIS is written in *Nextflow* (Di Tommaso et al., 2017), a high-level data-science language used to implement bioinformatics workflows and geared toward increasing portability of these workflows across different hardware and scheduling systems. The pipeline is fully open-source and licensed under GPLv3; the technical details are laid out in the “Methods” sections. Briefly, the pipeline takes a folder of images as an input, splits the total image set into batches of equal size, and performs semantic segmentation on the images using a model of choice (see section Advanced Models for Differentiated Plant Area Classification of Individual Leaves). The semantic segmentation output is then used to extract morphometric and color index traits from each individual image (see section Extraction of Morphometric and Color Index Traits from Segmented Images). Quality control and exploratory analyses can be performed after the pipeline has finished by launching a bundled *Shiny* (RStudio, 2014) application. In addition to offering some straightforward visualization of the results, *Shiny* also provides an interface to merge metadata and ARADEEPOPSIS output for downstream analysis (Figure 4; Di Tommaso et al., 2017). ARADEEPOPSIS is built modularly, and trait extraction works independently of the model: that is, although our pipeline comes with the three described models for plant segmentation (models A, B, and C; Figure 2), it offers flexibility for experienced users, via the option to feed in custom *TensorFlow* models (e.g., created using DPP). Alternatively, users can choose to use only the trait extraction and visualization steps of the pipeline by providing their own segmentation masks along with the original images.

ARADEEPOPSIS is designed to segment plants from images that contain only a single plant, and we do not recommend usage on images containing multiple plants. In case the phenotyping system records whole or partial trays of plant pots, as is common for many automated phenotyping platforms, users will need to preprocess these images and divide them into sub-images, each containing a single individual. Unless explicitly stated otherwise, we use ARADEEPOPSIS to refer to the pipeline using model C (Figure 2) from here on.

Validation and Generalization to Public Phenotyping Data Sets and Comparison to Other Models

To validate the accuracy of our pipeline, we re-analyzed publicly available images and compared the output of ARADEEPOPSIS to the published analyses. This validation step served two purposes: first, we wished to confirm that the rosette features extracted by ARADEEPOPSIS segmentation were accurately reproducing published data. Second, we wanted to test whether our pipeline

would remain robust when using data generated on different phenotyping platforms, with different image recording systems, pot sizes and shapes, and background composition. We used images from three published studies (Baute et al., 2017; Capovilla et al., 2018; Barragan et al., 2019), generated on either the RAPA platform (Vasseur et al., 2018) at the Max Planck Institute for Developmental Biology in Tübingen, Germany (Capovilla et al., 2018; Barragan et al., 2019), or the Weighing, Imaging and Watering Machine platform (<https://www.psb.ugent.be/phenotyping-platforms>) at the VIB in Ghent, Belgium (Baute et al., 2017; Coppens et al., 2017), and re-analyzed them using ARADEEPOPSIS.

Correlation analysis of the published rosette area derived from the RAPA system and our own measurements resulted in R^2 values of 0.996 (Capovilla et al., 2018) and 0.979 (Barragan et al., 2019), respectively (Figures 5A and 5B). Despite this overall very strong correlation, we noticed individual images in which our segmentation disagreed with the published record. When inspecting the respective images and segmentation masks from both analyses, we discovered that the respective rosette segmentations in the original analysis were inaccurate or incomplete, and that ARADEEPOPSIS segmentation was in strong agreement with the actual rosette in the image, even when plants had strong aberrant phenotypes or grew on low-contrast background (Figures 5A and 5B). We also used the pretrained *vegetationSegmentation* model from DPP to segment these two data sets and noticed a markedly lower to between DPP-derived and published measurements (Supplemental Figure 1A) when compared with ARADEEPOPSIS.

Re-analysis of the Weighing, Imaging and Watering Machine data set (Baute et al., 2017) also showed high correlation between the published measurements and our analysis (Supplemental Figure 1B). Unfortunately, original segmentations were not available for these images, but when inspecting some of the outliers, we again noticed highly accurate segmentation by ARADEEPOPSIS relative to the actual image (Supplemental Figure 1A).

To further validate the ARADEEPOPSIS pipeline, we assessed its performance on the International Plant Phenotyping Network (IPPN) data set (<https://www.plant-phenotyping.org/data-sets-download>; Minervini et al., 2016), which contains diverse images of plants with a range of backgrounds and varying image qualities. Importantly, this data set was not part of the training data set used during transfer learning. We used the ground-truth annotation masks of the IPPN data set with ARADEEPOPSIS to extract the respective ground-truth traits. We found that the models provided by ARADEEPOPSIS generalized well to this data set, although we noticed that some of the models (B and C) occasionally segmented background into classes corresponding to anthocyanin-rich or senescent, or misclassified algae as leaf tissue (Supplemental Figure 2B). Such artifacts can be reduced by running ARADEEPOPSIS with multi-scale inference, which improves segmentation accuracy by scaling the input image by different scale factors before model prediction and taking the average over these multiple scales for classification of individual pixels. With multi-scale inference, the segmentations of green plant tissue (ARADEEPOPSIS model C) agreed with the ground-truth annotations available for the IPPN data set (Figure 5C) and

Downloaded from https://academic.oup.com/plcell/article/32/12/3674/6118590 by guest on 14 July 2021

correlated better than those obtained when running our pipeline with the pretrained DPP vegetation segmentation model (Supplemental Figure 3). Because multi-scale inference results in higher computational demand, single-scale inference is the default setting; users will need to choose based on the availability of computational resources.

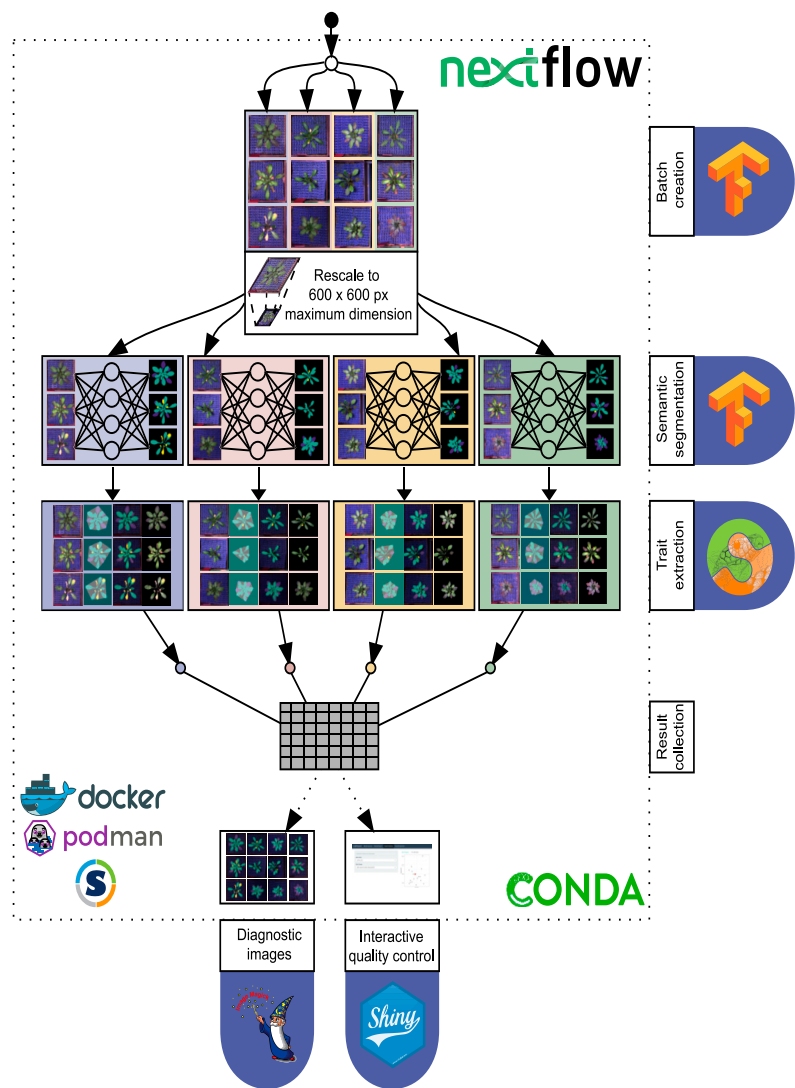


Figure 4. The ARADEEPOPSIS Pipeline. A folder containing image files is passed to the pipeline. The total number of images is split into batches of a fixed size (indicated by different background colors). Batches are processed in parallel: first, the segmentation is performed on each batch, then traits are extracted from the segmented images. The output from all batches is collected in one final table, the results table. In addition, the pipeline produces diagnostic images that show the segmentation results overlaid on the original image, color-coded segmentation masks, and background-subtracted plant rosettes. These diagnostic images can be explored in a *Shiny* (RStudio, 2014) app, which is launched at the end of an ARADEEPOPSIS run, such that users can visually inspect the segmentations and verify their accuracy.

3682 The Plant Cell

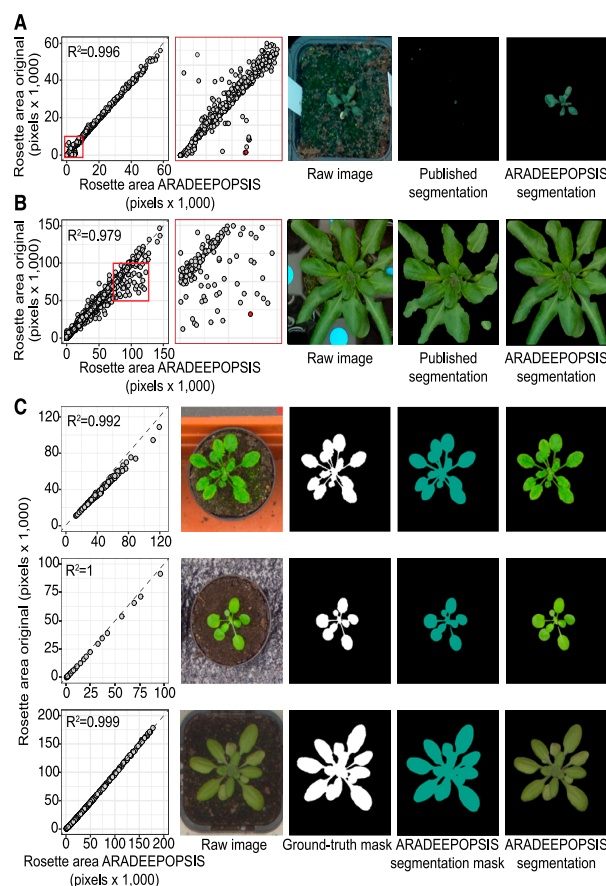


Figure 5. Validation of ARADEEPOPSIS Output Against Published Data.

(A) Validation against data from Capovilla et al. (2018). The left-most panel shows the correlation between values produced by ARADEEPOPSIS against published data; the boxed area is magnified in the second panel to highlight disagreements. The third panel shows the original image corresponding to the red data point in panel two. Panel four shows the original segmentation, panel five the segmentation by ARADEEPOPSIS.

(B) Validation against Barragan et al. (2019) with the same order as in (A).

(C) Validation against the A1 (top row), A2 (middle), and A4 (bottom) Arabidopsis image series (A1, A2, and A4) from the IPPN data set (Minervini et al., 2016). The ground-truth mask corresponds to the manually annotated foreground mask provided by IPPN.

Validation by GWA Studies

Validation of Color Index Traits

Often, the ultimate purpose of phenotypic measurements is to relate them to genetic data to understand the underlying genetic determinants. We therefore wanted to test whether ARADEEPOPSIS provided sufficiently high-quality data to identify genetic associations in Arabidopsis by GWA studies, relying on the 1001 genomes variant information (The 1001 Genomes Consortium, 2016). Instead of morphometric traits such as size and growth rate, which are usually highly polygenic and therefore difficult to test, we

wondered if we could make use of additional layers of image information provided by ARADEEPOPSIS as a proxy for physiological parameters. We hypothesized that RGB color information of the plant, that is, the respective pixel intensities in the three color channels, would provide information on anthocyanin content and hence on the physiological state of the plant, as it was previously shown that tissue color highly correlates with anthocyanin content (Del Valle et al., 2018). Using the ARADEEPOPSIS segmentations, which accurately reflect the area covered by the rosette, we extracted color information by collecting data from the different RGB channels from the segmented object area, as explained above

(Figures 3B and 3C). We then used *LIMIX* (Lippert et al., 2014) to perform GWA analysis on the “chroma ratio” of the rosette area 37 d after sowing. The “chroma ratio” is calculated by dividing the mean intensities from the green channel by half of the sum of the mean intensities from the blue and red channels (Figure 3C). The chroma ratio is therefore inversely correlated with anthocyanin accumulation and thus increases when anthocyanin content is low. We observed strong associations with regions on chromosomes 1, 2, 4, and 5 (Figure 6A). When ranked by $-\log_{10}(p)$, the second-highest ranking Single Nucleotide Polymorphism (SNP) mapped to chromosome 4, with the closest annotated gene being *ANTHOCYANINLESS2* (*ANL2*; At4g00730). *ANL2* encodes a homeobox transcription factor that has been implicated in controlling anthocyanin accumulation in sub-epidermal tissues (Kubo et al., 1999). Mutants in *ANL2* accumulate less anthocyanins in sub-epidermal leaf tissue. In our data, accessions carrying the alternative allele near *ANL2* displayed a higher chroma ratio, indicative of lower anthocyanin accumulation (Figure 6B). This result demonstrates that ARADEEPOPSIS can noninvasively extract quantitative color index data from whole rosettes that can be applied to genetic association analysis.

Validation of Pixel-Level Classifications

Having validated our hypothesis that color profiles of the whole plant area can be used to extract informative traits, we asked if

ARADEEPOPSIS would allow the identification of genetic variants significantly associated with the relative amount of senescent or dead tissue during early development. We conducted a GWA analysis, using the relative amount of senescent tissue at 25 d after sowing as the phenotypic trait. This time point is fairly early during the vegetative phase of plant development, and the relative senescent plant area is small compared to later stages (see Figure 2C). We identified several genomic regions that showed significant associations with this phenotype, the most striking of which was located on chromosome 4. The highest-ranking SNP in this region was located within the coding region of *ACCELERATED CELL DEATH 6* (*ACD6*; At4g14400; Figure 6C). *ACD6* has been extensively studied in Arabidopsis and was identified as being associated with vegetative growth, microbial resistance, and necrosis (Atwell et al., 2010; Todesco et al., 2010; Zhu et al., 2018). Our plants were grown on natural, nonsterilized soil from a field site near Zurich, Switzerland (Hu et al., 2018); it is therefore fair to assume that the microbial load and potential biotic stress levels were higher than in standard potting soil. We observed an ~8-fold difference in median relative senescent tissue per plant (Figure 6D) from ~0.5% of total segmented pixels in accessions carrying the reference allele to around 4% in those carrying the alternative allele. This result is in line with published studies, which highlighted the role of certain *ACD6* alleles in the early onset of senescence and necrosis in Arabidopsis (Todesco et al., 2010).

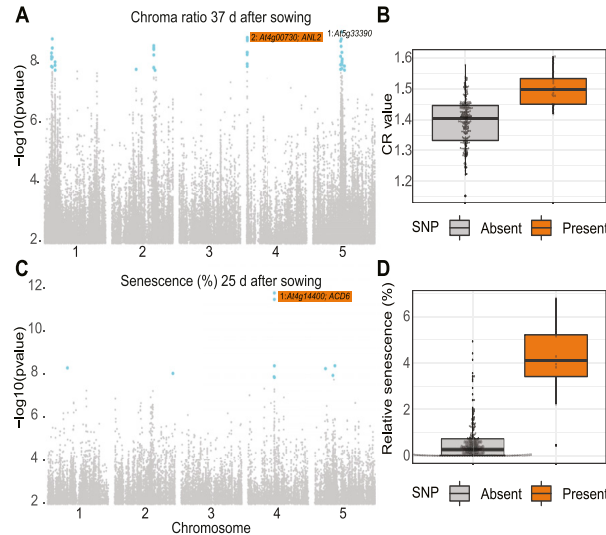


Figure 6. GWA Analyses based on ARADEEPOPSIS Output.

- (A) Results of GWA on the trait “chroma ratio” 37 d after sowing. The $\log_{10}$ -transformed p -value for each SNP is plotted (only SNPs with minor allele frequency >5% are shown). SNPs that are significant after Bonferroni correction are shown in blue. For the two highest-ranking SNPs, the closest gene is indicated. (B) Chroma ratio of plants 37 d after sowing, split by accessions carrying the reference and alternative alleles of the SNP close to At2g00730 (*ANTHOCYANINLESS2*, *ANL2*), highlighted in (A). (C) Same as (A) for the trait “relative senescence” 25 d after sowing. At4g14400 corresponds to *ACCELERATED CELL DEATH6* (*ACD6*, see text). (D) Relative senescence 25 d after sowing, split by accessions carrying the reference and alternative alleles of the SNP close to *ACD6*, highlighted in (C).

3684 The Plant Cell

Analysis of Other Plant Species

While our specific goal was to faithfully segment *Arabidopsis* rosettes, we wanted to see whether our method would generalize to other plant species with similar overall morphology. We therefore tested ARADEEPOPSIS on top-view images of pennycress (*Thlaspi arvense*), another member of the Brassicaceae, which has a similar leaf shape and phyllotaxis as *Arabidopsis* but does not form a flat rosette on the ground. Without training a model on images of *T. arvense*, that is, by using the model trained on *Arabidopsis*, segmentation masks accurately matched the original plant (Figure 7A), showing that ARADEEPOPSIS is robust against variations in plant morphology, as long as these do not deviate from the general architecture that was used in the training set. We also ran the pipeline on images from the IPPN data set (https://www.plant-phenotyping.org/data-sets-download; Minervini et al., 2016) on tobacco (*Nicotiana tabacum*), a member of the Solanaceae, which is morphologically quite distinct from *Arabidopsis*. Segmentations were accurate for most plants and correlated well with the ground-truth masks from the IPPN data set (Figure 7B). We did observe some exceptions for larger plants with substantial shaded leaf areas (Figure 7B; Supplemental Figure 2C).

Discussion

Here, we present ARADEEPOPSIS, a versatile pipeline to extract phenotypic measurements from small or large sets of plant images

in an unsupervised manner. ARADEEPOPSIS can faithfully identify, segment, and measure plants from top-view images of rosettes, such as those of *Arabidopsis*, using deep learning methodology. Our tool is easy to use, runs on personal computers as well as high-performance computing environments, is robust against variations in overall plant appearance, image quality, and background composition, and generates tabular output, making unsupervised high-throughput image-analysis accessible to a broad community of plant researchers.

A Fast and Easy-to-Use Segmentation Tool

Our models were trained on sets of 240 (model A) and >1,000 (models B and C) images, respectively, with manually annotated 'ground-truth' segmentations, which required substantial but manageable manual labor. We believe that these investments were warranted, because ARADEEPOPSIS produced accurate segmentations and morphometric measurements that correlated very well with published data; occasional deviations from the originally reported data were in most cases in favor of our pipeline, due to more accurate segmentation by ARADEEPOPSIS. In addition to these pretrained models, ARADEEPOPSIS also allows the use of custom models or sets of presegmented masks for use with the pipeline trait extraction module. Besides being accurate and requiring limited labor, the method is also fast: the fully

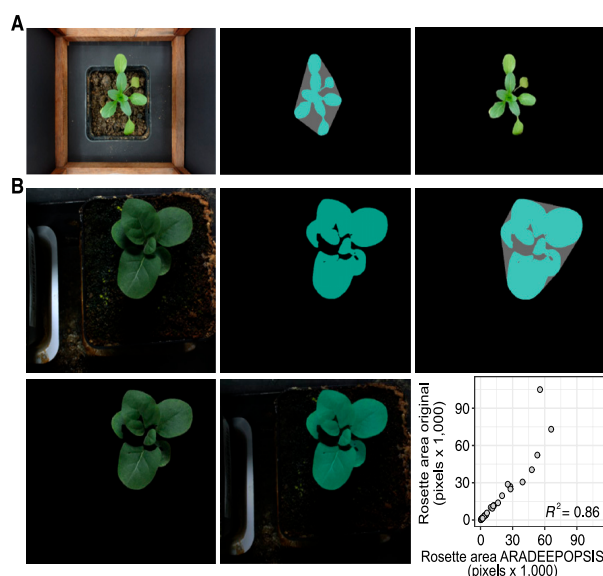


Figure 7. Segmentation of Other Plant Species.

(A) Segmentation of *Thlaspi arvense*. The left-most panel shows the original image of a *T. arvense* individual. The middle panel shows the segmented mask and convex hull area. The rightmost panel shows the plant accurately cropped from the original image using the mask generated by ARADEEPOPSIS. (B) Segmentation of a tobacco (*Nicotiana tabacum*) plant (https://www.plant-phenotyping.org/data-sets-download). From top left to center right, panels show the original image, the segmentation mask, the convex hull area, the segmentation, and the overlay of the segmentation and the original image of a representative plant. The scatterplot shows the correlation between ARADEEPOPSIS segmentation and the original ground-truth masks.

Downloaded from https://academic.oup.com/plcell/article/32/12/3674/6118590 by guest on 14 July 2021

automated analysis of 100 randomly selected images from our data set took 12 minutes on a laptop computer (24 Gbytes RAM, Intel i5-8250U 1.6 GHz). Depending on available resources such as memory and number of cores, the implementation in *Nextflow* enables straightforward parallelization, allowing for a significant increase in speed when deployed in high-performance computing environments. At the same time, *Nextflow* is agnostic to the computational infrastructure being used, making it straightforward to deploy ARADEEPOPSIS on any type of computer or cloud infrastructure, as long as it fulfills the minimum requirement of 6 Gbytes and 12 Gbytes of memory for single-scale and multi-scale inference, respectively. While training of the model greatly benefitted from the availability of graphics processing units (GPUs), image predictions can also be performed using central processing units in a time-efficient manner.

Accurate Determination of Plant Health State

ARADEEPOPSIS not only segments rosettes and extracts morphometric parameters, it also allows for the extraction of color channel information, which can be used to make assessments on the anthocyanin content and overall health status of the plant. We have shown that these results can be used, for example, for quantitative genetics approaches (Figure 6).

Besides a simple one-class model (model A) that can segment nonsenescent leaves, ARADEEPOPSIS includes two models that allow segmentation of anthocyanin-rich and/or senescent areas, depending on their color composition. As such, ARADEEPOPSIS is capable of reliably classifying senescent leaves and of distinguishing healthy, green leaves from stressed, anthocyanin-containing leaves. Our models may also be extended to additional classes, depending on the specific needs of researchers and the phenotypes of interest, by including the corresponding training sets.

Limitations and Shortcomings of the Proposed Models

Many of the images in our training data set had blue plastic meshes as background. This may raise the concern that the model might have learned to classify pixels belonging to the background rather than the plant of interest, that is, to segment the plant as “non-background,” which would render the tool unreliable when using it on images with a different background composition. However, by testing ARADEEPOPSIS on images acquired in other phenotyping environments using different backgrounds, and on images of potted plants with substantial algal growth surrounding the Arabidopsis plant of interest, we determined that segmentation of leaves remained accurate and was agnostic to the background (see Figure 5A). Ideally, ARADEEPOPSIS should be used with images that are homogeneous within one set regarding light intensity, exposure time, etc. Segmentation still works for images with various light intensities (Supplemental Figure 4), and plant size and geometry can still be analyzed, but quantification of color intensities based on the original image is no longer comparable in that case.

Our trained models have restrictions with regard to the angle at which the images are recorded and currently perform best for top-view images.

Although we found our models to generalize well to several publicly available data sets (Figure 5), there are also cases where the models misclassify parts of the background (Supplemental Figure 2). Such cases may effectively be avoided by re-training and fine-tuning the model, after adding a small number of annotated images representative of changed image parameters to the training data set. Such training set extensions may either be sourced from manual annotation or by selecting correctly segmented masks produced by the model. We will aim to further diversify the training set for a future iteration of model training and also encourage researchers to use and build upon the training code (https://github.com/phue/models/tree/aradeepopsis_manuscript/research/deeplab) and ground-truth annotations (<https://doi.org/10.5281/zenodo.3946393>) we deposited for that purpose.

When designing the pipeline, we placed a strong emphasis on the ability for the user to visually assess quality of segmentation at a glance, which we implemented using the *Shiny* framework (RStudio, 2014). This way, users can identify misclassifications early on. Because our pipeline calculates traits independently for each segmentation class, it can extract accurate information for some classes even when others are affected by misclassifications. We demonstrate this in Figure 5C, which shows that ARADEEPOPSIS measurements of class_norm_area correlate well with the area extracted from the IPPN CVPPP2017 annotations (<https://www.plant-phenotyping.org/data-sets-download>; Minervini et al., 2016), despite occasional misclassification of class_senescent, which partially occurred in the A1 image series of the data set (Supplemental Figure 2B).

Outlook and Perspective

Our study shows that transfer learning is a valuable approach to overcome the phenotyping bottleneck. The model was designed for the analysis of Arabidopsis rosettes but was also able to segment images from other species. We believe that providing a modular pipeline that can be used with different models provides valuable flexibility. For experimental conditions where our models do not perform well, when interest is in other species, or for analysis of other developmental traits than the ones implemented here, researchers can train their own models and simply integrate them into our pipeline to perform trait extraction. These adjustments will require expert annotation of the corresponding images but can be performed by an experienced researcher and trained helpers in a reasonable amount of time. We therefore anticipate that the straightforward usage of ARADEEPOPSIS for trait extraction with either the provided or custom models combined with direct visualization of the outputs will make semantic segmentation more easily accessible to plant scientists. Moreover, we would like to propose that pre-existing annotations from different research groups and various species be collected into a centralized repository that would serve as a powerful resource for phenomics in plant science and breeding.

ImageNet, the data set that was used to pretrain the baseline model we built upon, contains images of dogs, airplanes, bicycles,

3686 The Plant Cell

etc., but no detailed annotations of plant species. It is therefore remarkable that ImageNet pretrained models already enabled high accuracy for plant segmentation when re-training the last layer with a relatively small set of 240 manually annotated images. Ultimately, the adaptability displayed by ImageNet highlights the potential such publicly available databases and models hold for research and suggests that such resources should be exploited more frequently and more extensively.

Methods

Plant Growth and Image Acquisition

Plants (210 genotypes with 6 individuals each; 1260 plants total) were grown in a custom-built growth chamber (YORK Refrigeration), equipped with an environmental control unit (Johnson Controls) and a climate control unit (Incotec), on natural soil (Hu et al., 2018) in long-day conditions (16-h light [21°C], 8-h dark [16°C], 140 $\mu\text{E}/\text{m}^2\text{s}$) with 60% relative humidity and watered twice per week. As light sources, we used R-LMR-5001 light emitting diode panels (Rhenac); the spectrum is shown in Supplemental Figure 5. Seeds were surface-sterilized with chlorine gas for 1 h and sown on September 20, 2018, stratified for 4 d by cooling the chamber to 4°C, and from this point on plants were imaged two times per day. The images were acquired using an RGB camera (IDS uEye UI-548xRE-C; 5 MP) using an automated x-y-z sensor-to-plant system (Hesotech GmbH). Plants were monitored daily for emerging inflorescences, and flowering accessions were removed from the growth chamber to prevent emerging flowers from covering other plants in the top-down images. Plants that had not flowered by December 11, 2018 (12 weeks after release from stratification) were subjected to vernalization for 6 weeks: the chamber was cooled to 4°C (ramped over 8 h). During watering, the temperature was raised to 5°C. This treatment ended on January 21, 2019, after which the plants were exposed to the previous long-day growth conditions. In total 148,792 plant images were acquired and analyzed.

Training and Validation

We used the *Computer Vision Annotation Tool* (Sekachev et al., 2019) for manual image annotation; custom scripts were used to produce annotation masks from the XML output. The training data are available at <https://doi.org/10.5281/zenodo.3946393>. The publicly available *DeepLabV3+* (Chen et al., 2017; Chen et al., 2018) code was modified to enable model training on our own annotated training sets. The code used for training, as well as download links for our annotated training data sets, are available here: https://github.com/phue/models/tree/aradeepopsis_manuscript/research/deeplab.

For model evaluation, we split the annotated sets 80:20: 80% of the images were used to train the model and 20% for evaluation.

We used a transfer learning strategy by using a *DeepLabV3+* (based on the *xception65* architecture) model checkpoint (Chollet, 2017) that had been pretrained on the *ImageNet* data set (Deng et al., 2009) as a starting point. Training was then implemented in an asynchronous manner using between-graph replication on our in-house *Slurm* cluster, allowing the use of 16 Nvidia Tesla V100 GPUs across 4 compute nodes. We performed model training according to published protocols (Chen et al., 2017; Chen et al., 2018) with the following changes: to account for the number of GPUs, the training batch size was set to 128 and a base learning rate of 0.1 was applied and decayed according to a polynomial function after a burn-in period of 2000 training steps. Images were randomly cropped to 321 $\times$ 321 pixels, and training was stopped after 75,000 iterations. The models were exported to support single- or multiscale inference and are available at <https://doi.org/10.5281/zenodo.3946618>.

Implementation

Based on the trained models, we implemented an image analysis pipeline in *Nextflow* (Di Tommaso et al., 2017) that is available at <https://github.com/Gregor-Mendel-Institute/aradeepopsis>. The workflow is outlined in Figure 4. *Nextflow* allows external pipeline dependencies to be packaged in a *Docker* container, which we provide at <https://hub.docker.com/r/beckerlab/aradeepopsis/>. The container can be run using *Docker* (Merkel, 2014), *podman* (Containers Organisation, 2020), or *Singularity* (Kurtzer et al., 2017). Alternatively, dependencies can be automatically installed into a *Conda* (Anaconda Inc., 2020) environment.

The pipeline follows a scatter-gather strategy to scatter the input data into equally sized batches of arbitrary size, allowing for parallel processing, after which the results are again collected into a single output table. After the initial splitting, all images in one batch are first converted to TFrecord files using *TensorFlow* (Abadi et al., 2016). The records are then served to the trained *TensorFlow* model of choice in to generate segmentation masks, containing pixelwise classification.

In the next step, we extracted morphometric traits from the segmented images using the *regionprops* function of the Python library *scikit-image* (van der Walt et al., 2014) on a per-class basis. Additionally, color channel information was extracted in the form of mean pixel intensities for each of the channels in the original RGB image, within both the region determined by the segmentation mask and for each class. Based on the color channel means, we calculated chroma indices as follows (Del Valle et al., 2018):

$$\text{Chroma}_{\text{base}} = \frac{N_{\text{blue}} + N_{\text{red}}}{N_{\text{green}}}$$

$$\text{Chroma}_{\text{difference}} = \frac{N_{\text{blue}} + N_{\text{red}}}{2} - N_{\text{green}}$$

$$\text{Chroma}_{\text{ratio}} = \frac{N_{\text{green}}}{N_{\text{blue}} + N_{\text{red}}}$$

$$\text{BG}_{\text{ratio}} = \frac{N_{\text{blue}}}{N_{\text{green}}}$$

$$S_{\text{green}} = \frac{N_{\text{green}}}{N_{\text{red}} + N_{\text{green}} + N_{\text{blue}}}$$

The results from these measurements were then collected from all processed batches, ultimately resulting in a single table containing 78 traits, listed in the Supplemental Data Set. Based on the segmentation masks in combination with their corresponding input images, the pipeline also produces diagnostic images showing a color-coded mask, the cropped plant region, the convex hull, and an overlay image between the original and the mask.

These single-plant diagnostics are then, in an optional step, merged to produce summary diagnostics using *ImageMagick* (ImageMagick Development Team, 2020), or can be viewed in an interactive *Shiny* (RStudio, 2014) application specifically designed for the purpose of quality control of ARADEEPOPSIS output, allowing for fine-grained inspection of segmentation quality, pixel classification, correlations between traits, as well as time-resolved data visualization if appropriate metadata is provided.

GWA Studies

We performed GWA analysis on the traits produced by ARADEEPOPSIS using *LIMIX* (Lippert et al., 2014). We used the average of each trait per accession per d and performed GWA analysis by fitting a linear mixed model using *LIMIX*. To associate phenotype and single nucleotide polymorphisms, we used the 1135 genotype SNP matrix and the corresponding

kinship matrix, subset to those accessions for which we had trait information. We screened the results for interesting trait-date combinations and followed these using *Arabidopsis* (*Arabidopsis thaliana*)-specific R tools developed in-house (<https://github.com/Gregor-Mendel-Institute/gwaR>). The analysis is detailed in the Supplemental File.

Supplemental Material

Supplemental Figure 1. Validation of morphometric measurements against published data.

Supplemental Figure 2. Segmentations produced by ARADEEPOPSIS model C.

Supplemental Figure 3. Comparison of ARADEEPOPSIS to a dedicated small-scale model.

Supplemental Figure 4. ARADEEPOPSIS is tolerant to changes in image parameters.

Supplemental Figure 5. Light spectrum applied during the phenotyping experiment.

Supplemental Data Set. List of traits resulting from the ARADEEPOPSIS pipeline.

Supplemental File. R-scripts used for the analysis of GWAS results.

https://syncandshare.lrz.de/getlink/fILvKLBUsPKMrj4dLiGcklMa/araDeepopsis_supplement.html

AUTHOR CONTRIBUTIONS

P.H., N.S., and C.B. designed the research and wrote the article; P.H., N.S., and K.J. performed the research; P.H. wrote the pipeline; N.S. and K.J. performed the phenotyping experiment and generated the image data; I.B. provided additional image and segmentation data; P.H. and N.S. analyzed the data.

We thank J. Matthew Watson for critical reading of the manuscript and valuable comments. Karina Weiser Lobão helped in the manual image annotation. We thank Dario Galanti and Oliver Bosssdorf for providing images of *T. arvense*. Klaus Schlaeppli and Selma Cadot provided the field soil. We thank Núria Serra Serra and Jorge Isaac Rodriguez for testing the analysis pipeline. The computational results presented were obtained using the CLIP cluster at VBC. Erich Birngruber gave advice on distributed model training on a Slurm cluster. We thank Stijn Dhondt for providing access to the *Arabidopsis* images through the BrAPI endpoint of the PIPPA information system (<https://pippa.psb.ugent.be/>). *Arabidopsis* image acquisition was performed with support of the Plant Sciences Facility at Vienna BioCenter Core Facilities GmbH (VBCF), a member of the Vienna BioCenter (VBC), Austria. This work was supported by the EC | European Research Council (ERC; grant 716823, "FEAR-SAP" to C.B.), by the Austrian Academy of Sciences (Österreichische Akademie der Wissenschaften), and the Max-Planck-Gesellschaft (Max Planck Society).

Received April 24, 2020; revised September 28, 2020; accepted October 7, 2020; published October 9, 2020.

REFERENCES

Abadi, M., et al. (2016). TensorFlow: A system for large-scale machine learning. Conference Proceedings of the 12th USENIX

Symposium on Operating Systems Design and Implementation (OSDI '16); November 2–4, 2016; Savannah, GA.

Anaconda Inc (2020). Anaconda Software Distribution v. 2-2.4.0. Anaconda Documentation. <https://docs.anaconda.com/>

Arvidsson, S., Pérez-Rodríguez, P., and Mueller-Roeber, B. (2011). A growth phenotyping pipeline for *Arabidopsis thaliana* integrating image analysis and rosette area modeling for robust quantification of genotype effects. *New Phytol.* **191**: 895–907.

Atwell, S., et al. (2010). Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature* **465**: 627–631.

Barragan, C.A., et al. (2019). RPW8/HR repeats control NLR activation in *Arabidopsis thaliana*. *PLoS Genet.* **15**: e1008313.

Baute, J., Polyn, S., De Block, J., Blomme, J., Van Lijsebettens, M., and Inzé, D. (2017). F-Box protein FBX92 affects leaf size in *Arabidopsis thaliana*. *Plant Cell Physiol.* **58**: 962–975.

Capovilla, G., Delhomme, N., Collani, S., Shutava, I., Bezrukov, I., Symeonidis, E., de Francisco Amorim, M., Laubinger, S., and Schmid, M. (2018). PORCUPINE regulates development in response to temperature through alternative splicing. *Nat. Plants* **4**: 534–539.

Chen, L.-C., Papandreou, G., Schroff, F., and Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation. *ArXiv.org*: arXiv:1706.05587.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. ECCV 2018 – Computer Vision. In: Ferrari V., Hebert M., Sminchisescu C., and Weiss Y. (eds). Computer Vision – ECCV 2018. ECCV 2018. Lecture Notes in Computer Science, vol 11211. Springer, Cham, Switzerland. http://doi-org-443.webvpn.fjmu.edu.cn/10.1007/978-3-030-01234-2_49 European Conference on Computer Vision.

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. 2017 IEEE Conf Comp Vis Pattern Recogn (CVPR); Honolulu, HI; 2017; 1800–1807.

Containers Organisation (2020). Podman. <https://podman.io/>

Coppens, F., Wuyts, N., Inzé, D., and Dhondt, S. (2017). Unlocking the potential of plant phenotyping data through integration and data-driven approaches. *Curr. Opin. Syst. Biol.* **4**: 58–63.

Del Valle, J.C., Gallardo-López, A., Buide, M.L., Whittall, J.B., and Narbona, E. (2018). Digital photography provides a fast, reliable, and noninvasive method to estimate anthocyanin pigment concentration in reproductive and vegetative plant tissues. *Ecol. Evol.* **8**: 3064–3076.

Deng, J., Dong, W., Socher, R., Li, L., Li, K., and Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. IEEE Conference on Computer Vision and Pattern Recognition; Miami, FL; 2009; 248–255.

Di Tommaso, P., Chatzou, M., Floden, E.W., Barja, P.P., Palumbo, E., and Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* **35**: 316–319.

Furbank, R.T., and Tester, M. (2011). Phenomics—technologies to relieve the phenotyping bottleneck. *Trends Plant Sci.* **16**: 635–644.

Gehan, M.A., et al. (2017). PlantCV v2: Image analysis software for high-throughput plant phenotyping. *PeerJ* **5**: e4088.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016; 770–778.

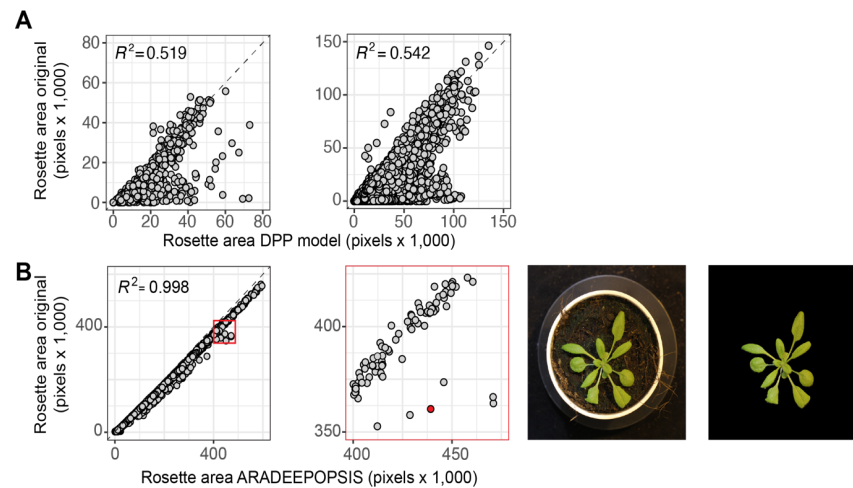
Hu, L., Robert, C.A.M., Cadot, S., Zhang, X., Ye, M., Li, B., Manzo, D., Chervet, N., Steinger, T., van der Heijden, M.G.A., Schlaeppli, K., and Erb, M. (2018). Root exudate metabolites drive plant-soil feedbacks on growth and defense by shaping the rhizosphere microbiota. *Nat. Commun.* **9**: 2738.

Krizhevsky, A., Sutskever, I., and Hinton, G.E. (2012). ImageNet classification with deep convolutional neural networks. *Nips* **12**: 1097–1105.

3688 The Plant Cell

- Kubo, H., Peeters, A.J., Aarts, M.G., Pereira, A., and Koorneef, M. (1999). ANTHOCYANINLESS2, a homeobox gene affecting anthocyanin distribution and root development in *Arabidopsis*. *Plant Cell* **11**: 1217–1226.
- Kurtzer, G.M., Sochat, V., and Bauer, M.W. (2017). Singularity: Scientific containers for mobility of compute. *PLoS One* **12**: e0177459.
- Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proc. IEEE* **86**: 2278–2324.
- Lippert, C., Casale, F.P., Rakitsch, B., and Stegle, O. (2014). LIMIX: Genetic analysis of multiple traits. *bioRxiv*: <https://doi.org/10.1101/003905>
- Merkel, D. (2014). Docker: Lightweight linux containers for consistent development and deployment. *Linux journal* **2014**: 2.
- Minervini, M., Fischbach, A., Scharr, H., and Tsaftaris, S.A. (2016). Finely-grained annotated datasets for image-based plant phenotyping. *Pattern Recognit. Lett.* **81**: 80–89.
- Minervini, M., Giuffrida, M.V., Perata, P., and Tsaftaris, S.A. (2017). Phenotiki: An open software and hardware platform for affordable and easy image-based phenotyping of rosette-shaped plants. *Plant J.* **90**: 204–216.
- Pound, M.P., Atkinson, J.A., Townsend, A.J., Wilson, M.H., Griffiths, M., Jackson, A.S., Bulat, A., Tzimiropoulos, G., Wells, D.M., Murchie, E.H., Pridmore, T.P., and French, A.P. (2017). Deep machine learning provides state-of-the-art performance in image-based plant phenotyping. *Gigascience* **6**: 1–10.
- Razavian, A.S., Azizpour, H., Sullivan, J., and Carlsson, S. (2014). CNN features off-the-shelf: an astounding baseline for recognition. 2014 IEEE Conf Comp Vis Pattern Recogn; 2016: 512–519.
- Rother, C., Kolmogorov, V., and Blake, A. (2004). "GrabCut": Interactive foreground extraction using iterated graph cuts. In ACM SIGGRAPH 2004 Papers. (Los Angeles, California: Association for Computing Machinery), pp. 309–314.
- RStudio (2014). Shiny: Easy web applications in R. R package. <https://rstudio.com/products/shiny/>
- Rumelhart, D.E., Hinton, G.E., and Williams, R.J. (1986). Learning representations by back-propagating errors. *Nature* **323**: 533–536.
- Schindelin, J., et al. (2012). Fiji: An open-source platform for biological-image analysis. *Nat. Methods* **9**: 676–682.
- Schneider, C.A., Rasband, W.S., and Eliceiri, K.W. (2012). NIH Image to ImageJ: 25 Years of image analysis. *Nat. Methods* **9**: 671–675.
- Sekachev, B., Zhavoronkov, A., and Manovich, N. (2019). Computer vision annotation tool: A universal approach to data annotation. software.intel.com.
- The 1001 Genomes Consortium. (2016). 1,135 Genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell* **166**: 481–491.
- Todesco, M., et al. (2010). Natural allelic variation underlying a major fitness trade-off in *Arabidopsis thaliana*. *Nature* **465**: 632–636.
- Ubbens, J., Cieslak, M., Prusinkiewicz, P., and Stavness, I. (2018). The use of plant models in deep learning: An application to leaf counting in rosette plants. *Plant Methods* **14**: 6.
- Ubbens, J.R., and Stavness, I. (2017). Deep Plant Phenomics: A deep learning platform for complex plant phenotyping tasks. *Front Plant Sci* **8**: 1190.
- van der Walt, S., Schonberger, J.L., Nunez-Iglesias, J., Boulogne, F., Warner, J.D., Yager, N., Gouillart, E., and Yu, T.; scikit-image contributors. (2014). scikit-image: image processing in Python. *PeerJ* **2**: e453.
- Vasseur, F., Bresson, J., Wang, G., Schwab, R., and Weigel, D. (2018). Image-based methods for phenotyping growth dynamics and fitness components in *Arabidopsis thaliana*. *Plant Methods* **14**: 63.
- Wang, C., Li, X., Caragea, D., Bheemanahallia, R., and Jagadish, S.V.K. (2020). Root anatomy based on root cross-section image analysis with deep learning. *Comp Electr Agricult* **175**: 105549.
- Yasrab, R., Atkinson, J.A., Wells, D.M., French, A.P., Pridmore, T.P., and Pound, M.P. (2019). RootNav 2.0: Deep learning for automatic navigation of complex plant root architectures. *Gigascience* **8**: 8.
- Zhu, W., et al. (2018). Modulation of ACD6 dependent hyperimmunity by natural alleles of an *Arabidopsis thaliana* NLR resistance gene. *PLoS Genet.* **14**: e1007628.

Downloaded from <https://academic.oup.com/plcell/article/32/12/3674/6118590> by guest on 14 July 2021

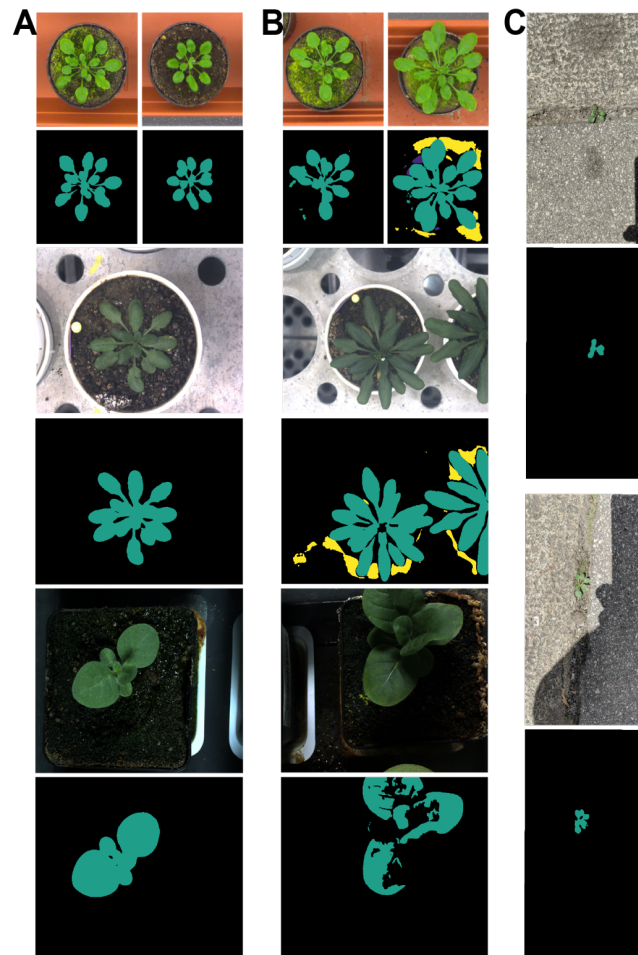


Supplemental Figure 1. Validation of morphometric measurements against published data. (Supports Figure 5).

(A) Correlation analysis of rosette measurements by the Deep Plant Phenomics model to the published data in Capovilla et al. (2018; left panel) and Barragan et al. (2018; right panel).

(B) The leftmost panel shows the correlation between values produced by ARADEEPOPSIS against published data (Baute et al. 2017). The second panel is a magnification of the boxed area in the first panel, highlighting disagreeing measurements. The third panel shows the original image of the red data point in panel two. Panel four shows the segmentation by ARADEEPOPSIS which is in strong agreement with the original image, suggesting inferior segmentation in the published analysis.

Supplemental Data, Hüther et al. (2020). The Plant Cell 10.1105/tpc.20.00318



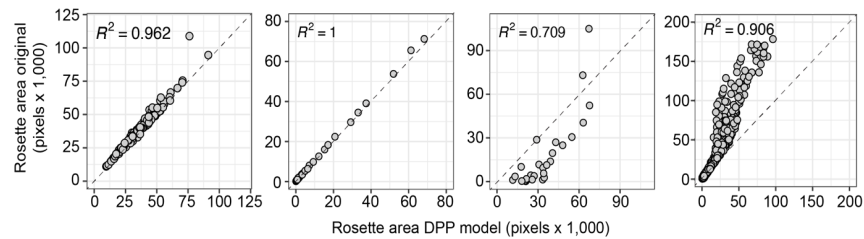
Supplemental Figure 2. Segmentations produced by ARADEEPOPSIS model C. (Supports Figure 5).

(A) Segmentations of plants selected from IPPN CVPPP2017 dataset (top images showing *Arabidopsis thaliana* and bottom image showing *Nicotiana tabacum*) and INRA Phenopsis dataset

(<http://bioweb.supagro.inra.fr/phenopsis/Accueil.php?lang=En>) (middle image showing *Arabidopsis thaliana*).

(B) Failure modes. Images were selected from the same datasets as in panel A. The model misclassifies algae as leaf tissue (top left) and suffers from misclassification of background trays in proximity of the plant (top left and middle image). It also has difficulties in correctly segmenting plants occluded by shadows (bottom).

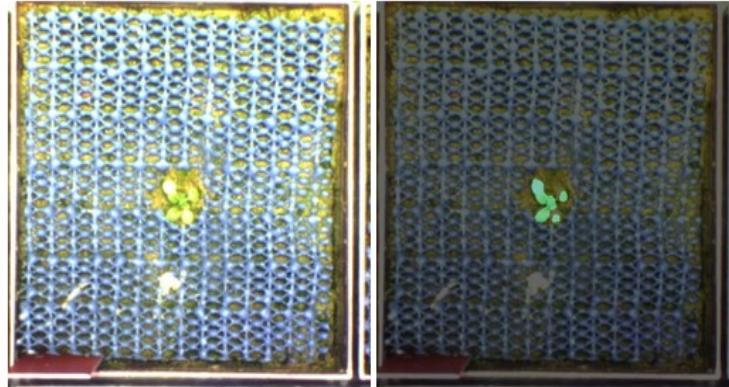
(C) Segmentation of pictures taken outdoors using a smartphone camera.



Supplemental Figure 3. Comparison of ARADEEPSIS to a dedicated small-scale model. (Supports Figure 5C).

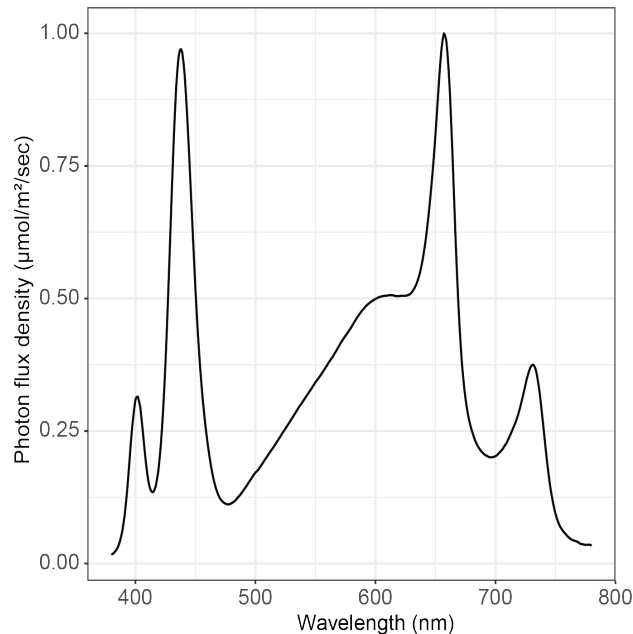
Correlation analysis of rosette measurements by the Deep Plant Phenomics (DPP) model to the ground-truth masks from the IPPN dataset (Minervini et al. 2016). From left to right, panels show correlations using the A1, A2, A3, and A4 image series of the IPPN datasets. DPP had been trained on the A1 and A2 series. A3 consists of images of *Nicotiana tabacum*; A4 was unseen to DPP. Dashed lines represent theoretical equivalent measurements between the model and the ground truth; note the strong skew of the data for A4.

Supplemental Data, Hüther et al. (2020). The Plant Cell 10.1105/tpc.20.00318



Supplemental Figure 4. ARADEEPOPSIS is tolerant to changes in image parameters. (Supports Figure 5).

Left: Original image, overexposed due to a technical problem with the light panel. Right: Segmented mask (green) overlayed with the original image.



Supplemental Figure 5. Light spectrum applied during the phenotyping experiment.

The spectrum generated by the R-LMR-5001 LED panels (Rhenac) that were used in the phenotyping assay, measured using a MK350S Premium handheld spectrometer (UPRtek).

Supplemental References

- Barragan, C.A., Wu, R., Kim, S.T., Xi, W., Habring, A., Hagmann, J., Van de Weyer, A.L., Zaidem, M., Ho, W.W.H., Wang, G., Bezrukov, I., Weigel, D., and Chae, E. (2019).** RPW8/HR repeats control NLR activation in *Arabidopsis thaliana*. *PLoS Genet* **15**, e1008313.
- Baute, J., Polyn, S., De Block, J., Blomme, J., Van Lijsebettens, M., and Inze, D. (2017).** F-Box protein FBX92 affects leaf size in *Arabidopsis thaliana*. *Plant Cell Physiol* **58**, 962-975.
- Capovilla, G., Delhomme, N., Collani, S., Shutava, I., Bezrukov, I., Symeonidi, E., de Francisco Amorim, M., Laubinger, S., and Schmid, M. (2018).** PORCUPINE regulates development in response to temperature through alternative splicing. *Nat Plants* **4**, 534-539.
- Minervini, M., Fischbach, A., Scharr, H., and Tsafaris, S.A. (2016).** Finely-grained annotated datasets for image-based plant phenotyping. *Pattern Recogn Lett* **81**, 8089.

6 Discussion

Chapter 2 presents a population-scale approach to detect differential methylation that is tailored to the context-specific methylation patterns found in plants. It showed that Methyl-Score can identify genomic regions displaying differential methylation from population-scale bisulfite sequencing data while avoiding assumptions about sample structure. Average methylation rates in identified regions hint at a profound interdependence between climatic conditions in natural habitats of *A. thaliana* and DNA methylation. Moreover, when using context-specific methylation rates for natural accessions as the phenotype vector in large-scale genome wide association studies (GWAS), in addition to previously described association at the CMT2 locus with CHH methylation (Sasaki et al. 2019), novel *trans*-acting associations pinpointing the MSI1 and miR823A loci could be identified that appear to modulate CHG methylation levels in certain regions of the genome. Intriguingly, a more recent study (Sasaki et al. 2022) confirms and further elucidates both loci we describe in CHG, using a conditional statistical framework that accounts for covariance between CHG and CHH methylation. In contrast to the genome-wide approach used in Chapter 2, the analysis in said study is restricted to transposable elements (and subclasses thereof) and elegantly shows that a total of five associations with putative regulators of CHG methylation are only found when correcting for CHH methylation, but not with univariate GWAS as it is used in Chapter 2. This indicates that our unsupervised classifier to segment context-specific DMRs is able to at least partially resolve the epistatic relationship between CHG and CHH methylation. The polymorphism found at the miR823A locus indicates a natural variant of a microRNA that is a known regulator of CMT3 expression. Its implications in embryonal development in *A. thaliana* were studied in Chapter 3. It was shown that the post-translational control that miR823A exerts on CMT3 transcript levels during embryogenesis is closely tied to the cell cycle and precludes the ectopic accumulation of CHG methylation on thousands of genes. However, the exact allelic effects and consequences of the natural variants identified in 2 on the phenotypic outcome remain to be studied in detail.

Furthermore, Chapter 2 shows the applicability of the methodology in common wildtype-mutant experimental designs, which is further exemplified by Chapter 4, in which the method was used for integrative analysis of DNA methylation, gene expression, and activity of transposable elements. The study shows that alterations in cellular levels of the methylgroup donor S-Adenosylmethionine (SAM) lead to measurable changes in DNA methylation as well as H3K9 methylation, and contributes to the understanding of the role of nitric oxide (NO) in epigenetic regulation of gene expression.

Chapter 5 of this thesis presents a robust approach for extracting phenotypic traits from top-view plant images at scale. Three trained CNN models for semantic segmentation and classification of several leaf states were obtained through transfer learning. The performance of these models could be successfully validated on RGB images from various sources, indicating that a well-generalising model was obtained from a training and validation set that was built from a limited number of annotated images containing plant

rosettes. The study showed the applicability of such methodology to a fairly large amount of 148.000 images, faithful and accurate extraction of morphometric and color-based traits, as well as quantification of growth, anthocyanin accumulation, and leaf senescence, spanning the entire vegetative growth phase across 210 *A. thaliana* accessions. The ability to successfully repurpose networks that were pre-trained on rather generic datasets such as ImageNet for specialised classification tasks in plant phenotyping clearly highlights the potential of such resources for future advances in the field of phenomics.

The specific implementation shown in Chapter 5 is restricted to top-view RGB images, which in turn constrains the approach to a subset of plant species that show the approximately 2-dimensional rosette-shaped vegetative growth pattern as is exemplified for *A. thaliana*. Transferring the approach to other classification tasks for different types of image data, such as side-view images, microscopy, histology, or herbaria images is well conceivable but would require training on novel annotations of the desired features. Often, however, phenotypic traits of interest are not limited to the quantification of vegetative growth but include photosynthetic activity, proxies of fitness such as yield, but also herbivore and pathogen infestation. While some of these traits are accessible through classical image based phenotyping, others are not and require more sophisticated techniques for data acquisition, for example light detection and ranging (LiDAR) that yields 3-dimensional point-cloud data, enabling measurement of various allometric parameters. While CNNs have proven useful for this type of data as well (Gong et al. 2021; Ghahremani et al. 2021), a transfer learning strategy is more difficult to implement because of the scarcity of publicly available datasets and pre-trained networks for that purpose, thus greatly increasing the necessary amounts of annotated training data when a model has to be designed and trained with random weight initialisation. In many ways, automated high-throughput plant phenotyping thus remains challenging in plant research and also breeding.

To conclude, the methodological aspects outlined in this thesis represent valuable contributions for population-scale studies in plant research. For instance, the combination of high-throughput plant phenotyping and methylome analysis enables epigenome wide association studies (EWAS) to further disentangle the interaction between phenotype and epigenotype, and will allow to further study and quantify the potential of epigenetic phenomena in shaping plant traits, their adaptive value and mechanistic detail.

As mentioned in the beginning, the astounding pace at which genomic data is produced from miniscule amounts of genetic material using next-generation sequencing has created an imbalance with regard to the rate at which phenotypic measurements can be obtained. Recent developments in long-read sequencing and ambitious endeavours such as the Earth BioGenome project (Lewin et al. 2018) suggest a rapidly growing repertoire of chromosome-level assemblies for a broad variety of species and ecotypes thereof. This will certainly open up exciting new avenues for genomics research within few years to come, but will at the same time exacerbate the persisting phenotyping bottleneck. In the context of plant breeding in particular, with increasing world population, occurrence of pests, and climate change as the major challenges of our time, there is an urge to find crop varieties with improved yield, nutrient content, enhanced drought tolerance or higher resilience - better sooner than later. Understanding the (epi-)genetic basis of an observed phenotype by drawing connections to available (epi)genomic information can greatly accelerate this process.

References

- 1001 Genomes Consortium (2016). ‘1,135 Genomes Reveal the Global Pattern of Polymorphism in *Arabidopsis thaliana*’. en. In: *Cell* 166.2, pp. 481–491.
- Abadi, Martín, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mane, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viegas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu and Xiaoqiang Zheng (2016). ‘TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems’. In: arXiv: 1603.04467 [cs.DC].
- Allfrey, V G, R Faulkner and A E Mirsky (1964). ‘Acetylation and Methylation of Histones and their Possible Role in the Regulation of RNA Synthesis’. In: *Proc. Natl. Acad. Sci. U. S. A.* 51.5, pp. 786–794.
- Arabidopsis Genome Initiative (2000). ‘Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*’. en. In: *Nature* 408.6814, pp. 796–815.
- Avery, Oswald T, Colin M MacLeod and Maclyn McCarty (1944). ‘Studies on the chemical nature of the substance inducing transformation of pneumococcal types induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III’. In: *J. Exp. Med.* 79.2, pp. 137–158.
- Ba, Jimmy Lei, Jamie Ryan Kiros and Geoffrey E Hinton (2016). ‘Layer Normalization’. In: arXiv: 1607.06450 [stat.ML].
- Bar, Yaniv, Idit Diamant, Lior Wolf, Sivan Lieberman, Eli Konen and Hayit Greenspan (2015). ‘Chest pathology detection using deep learning with non-medical training’. In: *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, pp. 294–297.
- Bednar, Jan, Rachel A Horowitz, Sergei A Grigoryev, Lenny M Carruthers, Jeffrey C Hansen, Abraham J Koster and Christopher L Woodcock (1998). ‘Nucleosomes, linker DNA, and linker histone form a unique structural motif that directs the higher-order folding and compaction of chromatin’. en. In: *Proc. Natl. Acad. Sci. U. S. A.* 95.24, pp. 14173–14178.
- Bernstein, Bradley E, Alexander Meissner and Eric S Lander (2007). ‘The mammalian epigenome’. en. In: *Cell* 128.4, pp. 669–681.
- Beucher, Serge and Christian Lantuéjoul (1979). ‘Use of watersheds in contour detection’. In: *Proceedings of the International Workshop on Image Processing*.
- Bewick, Adam J and Robert J Schmitz (2017). ‘Gene body DNA methylation in plants’. en. In: *Curr. Opin. Plant Biol.* 36, pp. 103–110.
- Bird, A P and E M Southern (1978). ‘Use of restriction enzymes to study eukaryotic DNA methylation: I. The methylation pattern in ribosomal DNA from *Xenopus laevis*’. en. In: *J. Mol. Biol.* 118.1, pp. 27–47.

- Bird, Adrian (2002). ‘DNA methylation patterns and epigenetic memory’. en. In: *Genes Dev.* 16.1, pp. 6–21.
- Borges, Filipe, Mark T A Donoghue, Chantal LeBlanc, Emily E Wear, Milos Tanurdžić, Benjamin Berube, Ashley Brooks, William F Thompson, Linda Hanley-Bowdoin and Robert A Martienssen (2021). ‘Loss of Small-RNA-Directed DNA Methylation in the Plant Cell Cycle Promotes Germline Reprogramming and Somaclonal Variation’. en. In: *Curr. Biol.* 31.3, 591–600.e4.
- Bradski, Gary (2000). ‘The OpenCV library’. In: *Dr Dobb’s J. Software Tools* 25, pp. 120–125.
- Braguy, Justine, Meray Ramazanov, Silvio Giancola, Muhammad Jamil, Boubacar A Kountche, Randa Zarban, Abrar Felemban, Jian You Wang, Pei-Yu Lin, Imran Haider, Matias Zurbriggen, Bernard Ghanem and Salim Al-Babili (2021). ‘SeedQuant: A deep learning-based tool for assessing stimulant and inhibitor activity on root parasitic seeds’. en. In: *Plant Physiol.*
- Brink, R A (1956). ‘A Genetic Change Associated with the R Locus in Maize Which Is Directed and Potentially Reversible’. en. In: *Genetics* 41.6, pp. 872–889.
- Cao, Xiaofeng, Werner Aufsatz, Daniel Zilberman, M Florian Mette, Michael S Huang, Marjori Matzke and Steven E Jacobsen (2003). ‘Role of the DRM and CMT3 methyltransferases in RNA-directed DNA methylation’. en. In: *Curr. Biol.* 13.24, pp. 2212–2217.
- Caselles, Vicent, Ron Kimmel and Guillermo Sapiro (1997). ‘Geodesic Active Contours’. In: *Int. J. Comput. Vis.* 22.1, pp. 61–79.
- Chan, Tony and Luminita Vese (1999). ‘An Active Contour Model without Edges’. In: *Scale-Space Theories in Computer Vision*. Springer Berlin Heidelberg, pp. 141–151.
- Chandler, Vicki L (2007). ‘Paramutation: from maize to mice’. en. In: *Cell* 128.4, pp. 641–645.
- Cheng, Shifeng, Michael Melkonian, Stephen A Smith, Samuel Brockington, John M Archibald, Pierre-Marc Delaux, Fay-Wei Li, Barbara Melkonian, Evgeny V Mavrodiev, Wenjing Sun, Yuan Fu, Huanming Yang, Douglas E Soltis, Sean W Graham, Pamela S Soltis, Xin Liu, Xun Xu and Gane Ka-Shu Wong (2018). ‘10KP: A phylodiverse genome sequencing plan’. en. In: *Gigascience* 7.3, pp. 1–9.
- Chollet, François et al. (2015). *Keras*. <https://keras.io>.
- Chollet, François (2016). ‘Xception: Deep Learning with Depthwise Separable Convolutions’. In: arXiv: 1610.02357 [cs.CV].
- Clevert, Djork-Arné, Thomas Unterthiner and Sepp Hochreiter (2015). ‘Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)’. In: arXiv: 1511.07289 [cs.LG].
- Cokus, Shawn J, Suhua Feng, Xiaoyu Zhang, Zugen Chen, Barry Merriman, Christian D Haudenschild, Sriharsa Pradhan, Stanley F Nelson, Matteo Pellegrini and Steven E Jacobsen (2008). ‘Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning’. en. In: *Nature* 452.7184, pp. 215–219.
- Cubas, P, C Vincent and E Coen (1999). ‘An epigenetic mutation responsible for natural variation in floral symmetry’. en. In: *Nature* 401.6749, pp. 157–161.
- Cuthbert, Graeme L, Sylvain Daujat, Andrew W Snowden, Hediye Erdjument-Bromage, Teruki Hagiwara, Michiyuki Yamada, Robert Schneider, Philip D Gregory, Paul Tempst, Andrew J Bannister and Tony Kouzarides (2004). ‘Histone deimination antagonizes arginine methylation’. en. In: *Cell* 118.5, pp. 545–553.

- Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li and Li Fei-Fei (2009). ‘ImageNet: A large-scale hierarchical image database’. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. Miami, FL: IEEE, pp. 248–255.
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee and Kristina Toutanova (2018). ‘BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding’. In: arXiv: 1810.04805 [cs.CL].
- Donahue, Jeff, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng and Trevor Darrell (2013). ‘DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition’. In: arXiv: 1310.1531 [cs.CV].
- Douarre, Clément, Richard Schielein, Carole Frindel, Stefan Gerth and David Rousseau (2018). ‘Transfer Learning from Synthetic Data Applied to Soil–Root Segmentation in X-Ray Tomography Images’. en. In: *Journal of Imaging* 4.5, p. 65.
- Du, Jiamu, Lianna M Johnson, Martin Groth, Suhua Feng, Christopher J Hale, Sisi Li, Ajay A Vashisht, James A Wohlschlegel, Dinshaw J Patel and Steven E Jacobsen (2014). ‘Mechanism of DNA methylation-directed histone methylation by KRYPTONITE’. en. In: *Mol. Cell* 55.3, pp. 495–504.
- Dunn, D B and J D Smith (1958). ‘The occurrence of 6-methylaminopurine in deoxyribonucleic acids’. en. In: *Biochem. J* 68.4, pp. 627–636.
- Ehrlich, M, M A Gama-Sosa, L H Carreira, L G Ljungdahl, K C Kuo and C W Gehrke (1985). ‘DNA methylation in thermophilic bacteria: N4-methylcytosine, 5-methylcytosine, and N6-methyladenine’. en. In: *Nucleic Acids Res.* 13.4, pp. 1399–1412.
- Everingham, Mark, Luc Gool, Christopher K Williams, John Winn and Andrew Zisserman (2010). ‘The Pascal Visual Object Classes (VOC) Challenge’. In: *Int. J. Comput. Vis.* 88.2, pp. 303–338.
- Fahlgren, Noah, Maximilian Feldman, Malia A Gehan, Melinda S Wilson, Christine Shyu, Douglas W Bryant, Steven T Hill, Colton J McEntee, Sankalpi N Warnasooriya, Indrajit Kumar, Tracy Ficor, Stephanie Turnipseed, Kerrigan B Gilbert, Thomas P Brutnell, James C Carrington, Todd C Mockler and Ivan Baxter (2015). ‘A Versatile Phenotyping System and Analytics Platform Reveals Diverse Temporal Responses to Water Availability in *Setaria*’. en. In: *Mol. Plant* 8.10, pp. 1520–1535.
- Feldmann, Mitchell J, Michael A Hardigan, Randi A Famula, Cindy M López, Amy Tabb, Glenn S Cole and Steven J Knapp (2020). ‘Multi-dimensional machine learning approaches for fruit shape phenotyping in strawberry’. en. In: *Gigascience* 9.5.
- Feng, Hao, Karen N Conneely and Hao Wu (2014). ‘A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data’. en. In: *Nucleic Acids Res.* 42.8, e69.
- Feng, Suhua, Steven E Jacobsen and Wolf Reik (2010). ‘Epigenetic reprogramming in plant and animal development’. en. In: *Science* 330.6004, pp. 622–627.
- Filippakopoulos, Panagis, Sarah Picaud, Maria Mangos, Tracy Keates, Jean-Philippe Lambert, Dalia Barsyte-Lovejoy, Ildiko Felletar, Rudolf Volkmer, Susanne Müller, Tony Pawson, Anne-Claude Gingras, Cheryl H Arrowsmith and Stefan Knapp (2012). ‘Histone recognition and large-scale structural analysis of the human bromodomain family’. en. In: *Cell* 149.1, pp. 214–231.
- Fischle, Wolfgang, Yanming Wang and C David Allis (2003). ‘Histone and chromatin cross-talk’. en. In: *Curr. Opin. Cell Biol.* 15.2, pp. 172–183.
- Frommer, M, L E McDonald, D S Millar, C M Collis, F Watt, G W Grigg, P L Molloy and C L Paul (1992). ‘A genomic sequencing protocol that yields a positive display of

- 5-methylcytosine residues in individual DNA strands'. en. In: *Proc. Natl. Acad. Sci. U. S. A.* 89.5, pp. 1827–1831.
- Gehan, Malia A, Noah Fahlgren, Arash Abbasi, Jeffrey C Berry, Steven T Callen, Leonardo Chavez, Andrew N Doust, Max J Feldman, Kerrigan B Gilbert, John G Hodge, J Steen Hoyer, Andy Lin, Suxing Liu, César Lizárraga, Argelia Lorence, Michael Miller, Eric Platon, Monica Tessman and Tony Sax (2017). 'PlantCV v2: Image analysis software for high-throughput plant phenotyping'. en. In: *PeerJ* 5, e4088.
- Ghahremani, Morteza, Kevin Williams, Fiona M K Corke, Bernard Tiddeman, Yonghuai Liu and John H Doonan (2021). 'Deep Segmentation of Point Clouds of Wheat'. en. In: *Front. Plant Sci.* 12, p. 608732.
- Glorot, Xavier and Yoshua Bengio (2010). 'Understanding the difficulty of training deep feedforward neural networks'. In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. Ed. by Yee Whye Teh and Mike Titterton. Vol. 9. Proceedings of Machine Learning Research. Chia Laguna Resort, Sardinia, Italy: PMLR, pp. 249–256.
- Goll, Mary Grace and Timothy H Bestor (2005). 'Eukaryotic cytosine methyltransferases'. en. In: *Annu. Rev. Biochem.* 74, pp. 481–514.
- Gong, Liang, Xiaofeng Du, Kai Zhu, Ke Lin, Qiaojun Lou, Zheng Yuan, Guoqiang Huang and Chengliang Liu (2021). 'Panicle-3D: Efficient Phenotyping Tool for Precise Semantic Segmentation of Rice Panicle Point Cloud'. en. In: *Plant Phenomics* 2021, p. 9838929.
- Grady, Leo (2006). 'Random walks for image segmentation'. en. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 28.11, pp. 1768–1783.
- Gruenbaum, Y, T Naveh-Many, H Cedar and A Razin (1981). 'Sequence specificity of methylation in higher plant DNA'. en. In: *Nature* 292.5826, pp. 860–862.
- Grunstein, M (1998). 'Yeast heterochromatin: regulation of its assembly and inheritance by histones'. en. In: *Cell* 93.3, pp. 325–328.
- Gulshan, Varun, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, Ramasamy Kim, Rajiv Raman, Philip C Nelson, Jessica L Mega and Dale R Webster (2016). 'Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs'. In: *JAMA* 316.22, pp. 2402–2410.
- Hagiwara, Teruki, Katsuhiko Nakashima, Hisashi Hirano, Tatsuo Senshu and Michiyuki Yamada (2002). 'Deimination of arginine residues in nucleophosmin/B23 and histones in HL-60 granulocytes'. en. In: *Biochem. Biophys. Res. Commun.* 290.3, pp. 979–983.
- Hamidinekoo, Azam, Gina A Garzón-Martínez, Morteza Ghahremani, Fiona M K Corke, Reyer Zwiggelaar, John H Doonan and Chuan Lu (2020). 'DeepPod: a convolutional neural network based quantification of fruit number in Arabidopsis'. en. In: *Gigascience* 9.3.
- Hassa, Paul O, Sandra S Haenni, Michael Elser and Michael O Hottiger (2006). 'Nuclear ADP-ribosylation reactions in mammalian cells: where are we today and where are we going?' en. In: *Microbiol. Mol. Biol. Rev.* 70.3, pp. 789–829.
- He, Kaiming, Xiangyu Zhang, Shaoqing Ren and Jian Sun (2015). 'Deep Residual Learning for Image Recognition'. In: arXiv: 1512.03385 [cs.CV].
- Hershey, A D and M Chase (1952). 'Independent functions of viral protein and nucleic acid in growth of bacteriophage'. en. In: *J. Gen. Physiol.* 36.1, pp. 39–56.
- Hinton, Geoffrey, Li Deng, Dong Yu, George E Dahl, Abdel-Rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath and Brian

- Kingsbury (2012a). ‘Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups’. In: *IEEE Signal Process. Mag.* 29.6, pp. 82–97.
- Hinton, Geoffrey E, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever and Ruslan R Salakhutdinov (2012b). ‘Improving neural networks by preventing co-adaptation of feature detectors’. In: arXiv: 1207.0580 [cs.NE].
- Hochreiter, S and J Schmidhuber (1997). ‘Long short-term memory’. en. In: *Neural Comput.* 9.8, pp. 1735–1780.
- Hoerl, Arthur E and Robert W Kennard (1970). ‘Ridge Regression: Biased Estimation for Nonorthogonal Problems’. In: *Technometrics* 12.1, pp. 55–67.
- Holliday, R and J E Pugh (1975). ‘DNA modification mechanisms and gene activity during development’. en. In: *Science* 187.4173, pp. 226–232.
- Hossain, Md Shakhawat, Taiji Kawakatsu, Kyung Do Kim, Ning Zhang, Cuong T Nguyen, Saad M Khan, Josef M Batek, Trupti Joshi, Jeremy Schmutz, Jane Grimwood, Robert J Schmitz, Dong Xu, Scott A Jackson, Joseph R Ecker and Gary Stacey (2017). ‘Divergent cytosine DNA methylation patterns in single-cell, soybean root hairs’. en. In: *New Phytol.* 214.2, pp. 808–819.
- Hubel, D H and T N Wiesel (1959). ‘Receptive fields of single neurones in the cat’s striate cortex’. en. In: *J. Physiol.* 148, pp. 574–591.
- Innes, Michael, Elliot Saba, Keno Fischer, Dhairya Gandhi, Marco Concetto Rudilosso, Neethu Mariya Joy, Tejan Karmali, Avik Pal and Viral Shah (2018). ‘Fashionable Modelling with Flux’. In: arXiv: 1811.01457 [cs.PL].
- Ioffe, Sergey and Christian Szegedy (2015). ‘Batch normalization: accelerating deep network training by reducing internal covariate shift’. In: *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*. ICML’15. Lille, France: JMLR.org, pp. 448–456.
- Jackson, James P, Anders M Lindroth, Xiaofeng Cao and Steven E Jacobsen (2002). ‘Control of CpNpG DNA methylation by the KRYPTONITE histone H3 methyltransferase’. en. In: *Nature* 416.6880, pp. 556–560.
- Jacobs, Steven A and Sepideh Khorasanizadeh (2002). ‘Structure of HP1 chromodomain bound to a lysine 9-methylated histone H3 tail’. en. In: *Science* 295.5562, pp. 2080–2083.
- Jang, Chuan-Wei, Yoichiro Shibata, Joshua Starmer, Della Yee and Terry Magnuson (2015). ‘Histone H3.3 maintains genome integrity during mammalian development’. en. In: *Genes Dev.* 29.13, pp. 1377–1392.
- Jenuwein, T and C D Allis (2001). ‘Translating the histone code’. en. In: *Science* 293.5532, pp. 1074–1080.
- Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Židek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Andrew J Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W Senior, Koray Kavukcuoglu, Pushmeet Kohli and Demis Hassabis (2021). ‘Highly accurate protein structure prediction with AlphaFold’. en. In: *Nature*.
- Kantz, Edward D, Saumya Tiwari, Jeramie D Watrous, Susan Cheng and Mohit Jain (2019). ‘Deep Neural Networks for Classification of LC-MS Spectral Peaks’. en. In: *Anal. Chem.* 91.19, pp. 12407–12413.

- Kawakatsu, Taiji, Shao-Shan Carol Huang, Florian Jupe, Eriko Sasaki, Robert J Schmitz, Mark A Urich, Rosa Castanon, Joseph R Nery, Cesar Barragan, Yupeng He, Huaming Chen, Manu Dubin, Cheng-Ruei Lee, Congmao Wang, Felix Bemm, Claude Becker, Ryan O’Neil, Ronan C O’Malley, Danjuma X Quarless, 1001 Genomes Consortium, Nicholas J Schork, Detlef Weigel, Magnus Nordborg and Joseph R Ecker (2016). ‘Epigenomic Diversity in a Global Collection of Arabidopsis thaliana Accessions’. en. In: *Cell* 166.2, pp. 492–505.
- Kelley, David R, Jasper Snoek and John L Rinn (2016). ‘Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks’. en. In: *Genome Res.* 26.7, pp. 990–999.
- Köhler, C and I Weinhofer-Molisch (2010). ‘Mechanisms and evolution of genomic imprinting in plants’. en. In: *Heredity* 105.1, pp. 57–63.
- Kornberg, R D and J O Thomas (1974). ‘Chromatin structure; oligomers of the histones’. en. In: *Science* 184.4139, pp. 865–868.
- Kouzarides, Tony (2007). ‘Chromatin Modifications and Their Function’. In: *Cell* 128.4, pp. 693–705.
- Krizhevsky, Alex, Ilya Sutskever and Geoffrey E Hinton (2012). ‘ImageNet Classification with Deep Convolutional Neural Networks’. In: *Advances in Neural Information Processing Systems*. Ed. by F Pereira, C J C Burges, L Bottou and K Q Weinberger. Vol. 25. Curran Associates, Inc.
- Krueger, Felix and Simon R Andrews (2011). ‘Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications’. en. In: *Bioinformatics* 27.11, pp. 1571–1572.
- Kurdistani, S K and M Grunstein (2003). ‘Histone acetylation and deacetylation in yeast’. en. In: *Nat. Rev. Mol. Cell Biol.* 4.4, pp. 276–284.
- Lachner, M, D O’Carroll, S Rea, K Mechtler and T Jenuwein (2001). ‘Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins’. en. In: *Nature* 410.6824, pp. 116–120.
- Law, Julie A and Steven E Jacobsen (2010). ‘Establishing, maintaining and modifying DNA methylation patterns in plants and animals’. en. In: *Nat. Rev. Genet.* 11.3, pp. 204–220.
- LeCun, Y, B Boser, J S Denker, D Henderson, R E Howard, W Hubbard and L D Jackel (1989). ‘Backpropagation Applied to Handwritten Zip Code Recognition’. In: *Neural Comput.* 1.4, pp. 541–551.
- Lecun, Y, L Bottou, Y Bengio and P Haffner (1998). ‘Gradient-based learning applied to document recognition’. In: *Proc. IEEE* 86.11, pp. 2278–2324.
- Lewin, Harris A, Gene E Robinson, W John Kress, William J Baker, Jonathan Coddington, Keith A Crandall, Richard Durbin, Scott V Edwards, Félix Forest, M Thomas P Gilbert, Melissa M Goldstein, Igor V Grigoriev, Kevin J Hackett, David Haussler, Erich D Jarvis, Warren E Johnson, Aristides Patrinos, Stephen Richards, Juan Carlos Castilla-Rubio, Marie-Anne van Sluys, Pamela S Soltis, Xun Xu, Huanming Yang and Guojie Zhang (2018). ‘Earth BioGenome Project: Sequencing life for the future of life’. en. In: *Proc. Natl. Acad. Sci. U. S. A.* 115.17, pp. 4325–4333.
- Lin, Tsung-Yi, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár and C Lawrence Zitnick (2014). ‘Microsoft COCO: Common Objects in Context’. In: *Computer Vision – ECCV 2014*. Springer International Publishing, pp. 740–755.

- Lin, Yingxin, Tung-Yu Wu, Sheng Wan, Jean Y H Yang, Wing H Wong and Y X Rachel Wang (2022). ‘scJoint integrates atlas-scale single-cell RNA-seq and ATAC-seq data with transfer learning’. en. In: *Nat. Biotechnol.*
- Lindroth, A M, X Cao, J P Jackson, D Zilberman, C M McCallum, S Henikoff and S E Jacobsen (2001). ‘Requirement of CHROMOMETHYLASE3 for maintenance of CpXpG methylation’. en. In: *Science* 292.5524, pp. 2077–2080.
- Lister, Ryan, Ronan C O’Malley, Julian Tonti-Filippini, Brian D Gregory, Charles C Berry, A Harvey Millar and Joseph R Ecker (2008). ‘Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis’. en. In: *Cell* 133.3, pp. 523–536.
- Litt, M D, M Simpson, M Gaszner, C D Allis and G Felsenfeld (2001). ‘Correlation between histone lysine methylation and developmental changes at the chicken beta-globin locus’. en. In: *Science* 293.5539, pp. 2453–2455.
- Lloyd, S (1982). ‘Least squares quantization in PCM’. In: *IEEE Trans. Inf. Theory* 28.2, pp. 129–137.
- Lu, Rui and Gang Greg Wang (2013). ‘Tudor: a versatile family of histone methylation ‘readers’’. en. In: *Trends Biochem. Sci.* 38.11, pp. 546–555.
- Luger, K, A W Mäder, R K Richmond, D F Sargent and T J Richmond (1997). ‘Crystal structure of the nucleosome core particle at 2.8 Å resolution’. en. In: *Nature* 389.6648, pp. 251–260.
- Ma, Junshui, Robert P Sheridan, Andy Liaw, George E Dahl and Vladimir Svetnik (2015). ‘Deep Neural Nets as a Method for Quantitative Structure–Activity Relationships’. In: *J. Chem. Inf. Model.* 55.2, pp. 263–274.
- MacAlpine, David M and Geneviève Almouzni (2013). ‘Chromatin and DNA replication’. en. In: *Cold Spring Harb. Perspect. Biol.* 5.8, a010207.
- Manolio, Teri A, Francis S Collins, Nancy J Cox, David B Goldstein, Lucia A Hindorff, David J Hunter, Mark I McCarthy, Erin M Ramos, Lon R Cardon, Aravinda Chakravarti, Judy H Cho, Alan E Guttmacher, Augustine Kong, Leonid Kruglyak, Elaine Mardis, Charles N Rotimi, Montgomery Slatkin, David Valle, Alice S Whittemore, Michael Boehnke, Andrew G Clark, Evan E Eichler, Greg Gibson, Jonathan L Haines, Trudy F C Mackay, Steven A McCarroll and Peter M Visscher (2009). ‘Finding the missing heritability of complex diseases’. en. In: *Nature* 461.7265, pp. 747–753.
- Márquez-Neila, Pablo, Luis Baumela and Luis Alvarez (2014). ‘A morphological approach to curvature-based evolution of curves and surfaces’. en. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 36.1, pp. 2–17.
- Mayr, A, G Klambauer, T Unterthiner et al. (2016). ‘DeepTox: toxicity prediction using deep learning’. In: *Frontiers in*.
- McCulloch, Warren S and Walter Pitts (1943). ‘A logical calculus of the ideas immanent in nervous activity’. en. In: *Bull. Math. Biophys.* 5.4, pp. 115–133.
- Meselson, M and F W Stahl (1958). ‘THE REPLICATION OF DNA IN ESCHERICHIA COLI’. en. In: *Proc. Natl. Acad. Sci. U. S. A.* 44.7, pp. 671–682.
- Meyer, P, I Niedenhof and M ten Lohuis (1994). ‘Evidence for cytosine methylation of non-symmetrical sequences in transgenic *Petunia hybrida*’. en. In: *EMBO J.* 13.9, pp. 2084–2088.
- Minsky, Marvin and Seymour A Papert (1969). *Perceptrons: An Introduction to Computational Geometry*. en. MIT Press.
- Miura, Kotaro, Masakazu Agetsuma, Hidemi Kitano, Atsushi Yoshimura, Makoto Matsuoka, Steven E Jacobsen and Motoyuki Ashikari (2009). ‘A metastable DWARF1 epigenetic

- mutant affecting plant stature in rice'. en. In: *Proc. Natl. Acad. Sci. U. S. A.* 106.27, pp. 11218–11223.
- Molaro, Antoine, Emily Hodges, Fang Fang, Qiang Song, W Richard McCombie, Gregory J Hannon and Andrew D Smith (2011). 'Sperm methylation profiles reveal features of epigenetic inheritance and evolution in primates'. en. In: *Cell* 146.6, pp. 1029–1041.
- Morrison, Ashby J and Xuotong Shen (2005). 'DNA repair in the context of chromatin'. en. In: *Cell Cycle* 4.4, pp. 568–571.
- Nair, Vinod and Geoffrey E Hinton (2010). 'Rectified linear units improve restricted boltzmann machines'. In: *Proceedings of the 27th International Conference on International Conference on Machine Learning*. ICML'10. Haifa, Israel: Omnipress, pp. 807–814.
- Nanney, D L (1958). 'EPIGENETIC CONTROL SYSTEMS'. en. In: *Proc. Natl. Acad. Sci. U. S. A.* 44.7, pp. 712–717.
- Nathan, Dafna, Kristin Ingvarsdottir, David E Sterner, Gwendolyn R Bylebyl, Milos Dokmanovic, Jean A Dorsey, Kelly A Whelan, Mihajlo Krsmanovic, William S Lane, Pamela B Meluh, Erica S Johnson and Shelley L Berger (2006). 'Histone sumoylation is a negative regulator in *Saccharomyces cerevisiae* and shows dynamic interplay with positive-acting histone modifications'. en. In: *Genes Dev.* 20.8, pp. 966–976.
- Niblack, Wayne (1985). *An introduction to digital image processing*. DNK: Strandberg Publishing Company.
- Nielsen, Peter R, Daniel Nietlispach, Helen R Mott, Juliana Callaghan, Andrew Bannister, Tony Kouzarides, Alexey G Murzin, Natalia V Murzina and Ernest D Laue (2002). 'Structure of the HP1 chromodomain bound to histone H3 methylated at lysine 9'. en. In: *Nature* 416.6876, pp. 103–107.
- Ning, Yong-Qiang, Na Liu, Ke-Ke Lan, Yin-Na Su, Lin Li, She Chen and Xin-Jian He (2020). 'DREAM complex suppresses DNA methylation maintenance genes and precludes DNA hypermethylation'. en. In: *Nat Plants* 6.8, pp. 942–956.
- Nowak, Scott J and Victor G Corces (2004). 'Phosphorylation of histone H3: a balancing act between chromosome condensation and transcriptional activation'. In: *Trends Genet.* 20.4, pp. 214–220.
- Olins, A L and D E Olins (1974). 'Spheroid chromatin units (v bodies)'. en. In: *Science* 183.4122, pp. 330–332.
- Olova, Nelly, Felix Krueger, Simon Andrews, David Oxley, Rebecca V Berrens, Miguel R Branco and Wolf Reik (2018). 'Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting DNA methylation data'. en. In: *Genome Biol.* 19.1, p. 33.
- Olsson, Ola, Melanie Karlsson, Anna S Persson, Henrik G Smith, Vidula Varadarajan, Johanna Yourstone and Martin Stjernman (2021). 'Efficient, automated, and robust pollen analysis using deep learning'. en. In: *Methods Ecol. Evol.* 2041-210X.13575.
- Ong-Abdullah, Meilina, Jared M Ordway, Nan Jiang, Siew-Eng Ooi, Sau-Yee Kok, Norashikin Sarpan, Nuraziyah Azimi, Ahmad Tarmizi Hashim, Zamzuri Ishak, Samsul Kamal Rosli, Fadila Ahmad Malike, Nor Azwani Abu Bakar, Marhalil Marjuni, Norziha Abdullah, Zulkifli Yaakub, Mohd Din Amiruddin, Rajanaidu Nookiah, Rajinder Singh, Eng-Ti Leslie Low, Kuang-Lim Chan, Norazah Azizi, Steven W Smith, Blaire Bacher, Muhammad A Budiman, Andrew Van Brunt, Corey Wischmeyer, Melissa Beil, Michael Hogan, Nathan Lakey, Chin-Ching Lim, Xaviar Arulandoo, Choo-Kien Wong, Chin-Nee Choo, Wei-Chee Wong, Yen-Yen Kwan, Sharifah Shahrul Rabiah Alwee, Ravigadevi

- Sambanthamurthi and Robert A Martienssen (2015). ‘Loss of Karma transposon methylation underlies the mantled somaclonal variant of oil palm’. en. In: *Nature* 525.7570, pp. 533–537.
- Osley, Mary Ann, Toyoko Tsukuda and Jac A Nickoloff (2007). ‘ATP-dependent chromatin remodeling factors and DNA damage repair’. en. In: *Mutat. Res.* 618.1-2, pp. 65–80.
- Otsu, Nobuyuki (1979). ‘A Threshold Selection Method from Gray-Level Histograms’. In: *IEEE Trans. Syst. Man Cybern.* 9.1, pp. 62–66.
- Owen, D J, P Ornaghi, J C Yang, N Lowe, P R Evans, P Ballario, D Neuhaus, P Filetici and A A Travers (2000). ‘The structural basis for the recognition of acetylated histone H4 by the bromodomain of histone acetyltransferase gcn5p’. en. In: *EMBO J.* 19.22, pp. 6141–6149.
- Park, Yongseok and Hao Wu (2016). ‘Differential methylation analysis for BS-seq data under general experimental design’. en. In: *Bioinformatics* 32.10, pp. 1446–1453.
- Paszke, Adam, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai and Soumith Chintala (2019). ‘PyTorch: An Imperative Style, High-Performance Deep Learning Library’. In: *Advances in Neural Information Processing Systems*. Ed. by H Wallach, H Larochelle, A Beygelzimer, F Alché-Buc, E Fox and R Garnett. Vol. 32. Curran Associates, Inc.
- Pedersen, Brent S, Kenneth Eyring, Subhajyoti De, Ivana V Yang and David A Schwartz (2014). ‘Fast and accurate alignment of long bisulfite-seq reads’. In: arXiv: 1401.1129 [q-bio.GN].
- Polder, Gerrit, Pieter M Blok, Hendrik A C de Villiers, Jan M van der Wolf and Jan Kamp (2019). ‘Potato Virus Y Detection in Seed Potatoes Using Deep Learning on Hyperspectral Images’. en. In: *Front. Plant Sci.* 10, p. 209.
- Poplin, Ryan, Pi-Chuan Chang, David Alexander, Scott Schwartz, Thomas Colthurst, Alexander Ku, Dan Newburger, Jojo Dijamco, Nam Nguyen, Pegah T Afshar, Sam S Gross, Lizzie Dorfman, Cory Y McLean and Mark A DePristo (2018). ‘A universal SNP and small-indel variant caller using deep neural networks’. en. In: *Nat. Biotechnol.* 36.10, pp. 983–987.
- Razavian, Ali Sharif, Hossein Azizpour, Josephine Sullivan and Stefan Carlsson (2014). ‘CNN Features off-the-shelf: an Astounding Baseline for Recognition’. In: arXiv: 1403.6382 [cs.CV].
- Richards, Eric J and Sarah C R Elgin (2002). ‘Epigenetic codes for heterochromatin formation and silencing: rounding up the usual suspects’. en. In: *Cell* 108.4, pp. 489–500.
- Riggs, A D (1975). ‘X inactivation, differentiation, and DNA methylation’. en. In: *Cytogenet. Cell Genet.* 14.1, pp. 9–25.
- Rosenblatt, F (1958). ‘The perceptron: A probabilistic model for information storage and organization in the brain’. In: *Psychol. Rev.* 65.6, pp. 386–408.
- Rother, Carsten, Vladimir Kolmogorov and Andrew Blake (2004). ‘GrabCut: interactive foreground extraction using iterated graph cuts’. In: *ACM SIGGRAPH 2004 Papers. SIGGRAPH ’04*. Los Angeles, California: Association for Computing Machinery, pp. 309–314.
- Rumelhart, David E, Geoffrey E Hinton and Ronald J Williams (1986). ‘Learning representations by back-propagating errors’. In: *Nature* 323.6088, pp. 533–536.

- Russakovsky, Olga, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C Berg and Li Fei-Fei (2015). ‘ImageNet Large Scale Visual Recognition Challenge’. In: *Int. J. Comput. Vis.* 115.3, pp. 211–252.
- Russo, Vincenzo E A, Robert A Martienssen and Arthur D Riggs (1996). *Epigenetic mechanisms of gene regulation*. Cold Spring Harbor Laboratory Press.
- Sabari, Benjamin R, Di Zhang, C David Allis and Yingming Zhao (2016). ‘Metabolic regulation of gene expression through histone acylations’. en. In: *Nat. Rev. Mol. Cell Biol.* 18.2, pp. 90–101.
- Salimans, Tim and Durk P Kingma (2016). ‘Weight Normalization: A Simple Reparameterization to Accelerate Training of Deep Neural Networks’. In: *Advances in Neural Information Processing Systems*. Ed. by D Lee, M Sugiyama, U Luxburg, I Guyon and R Garnett. Vol. 29. Curran Associates, Inc.
- Sankaran, Sindhuja, Lav R Khot, Carlos Zúñiga Espinoza, Sanaz Jarolmasjed, Vidyasagar R Sathuvalli, George J Vandemark, Phillip N Miklas, Arron H Carter, Michael O Pumphrey, N Richard Knowles and Mark J Pavak (2015). ‘Low-altitude, high-resolution aerial imaging systems for row and field crop phenotyping: A review’. In: *Eur. J. Agron.* 70, pp. 112–123.
- Sasaki, Eriko, Joanna Gunis, Ilka Reichardt-Gomez, Viktoria Nizhynska and Magnus Nordborg (2022). ‘Conditional GWAS of non-CG transposon methylation in *Arabidopsis thaliana* reveals major polymorphisms in five genes’. en.
- Sasaki, Eriko, Taiji Kawakatsu, Joseph R Ecker and Magnus Nordborg (2019). ‘Common alleles of CMT2 and NRPE1 are major determinants of CHH methylation variation in *Arabidopsis thaliana*’. en. In: *PLoS Genet.* 15.12, e1008492.
- Sauvola, J and M Pietikäinen (2000). ‘Adaptive document image binarization’. In: *Pattern Recognit.* 33.2, pp. 225–236.
- Saxe, Andrew M, James L McClelland and Surya Ganguli (2013). ‘Exact solutions to the nonlinear dynamics of learning in deep linear neural networks’. In: arXiv: 1312.6120 [cs.NE].
- Schindelin, Johannes, Ignacio Arganda-Carreras, Erwin Frise, Verena Kaynig, Mark Longair, Tobias Pietzsch, Stephan Preibisch, Curtis Rueden, Stephan Saalfeld, Benjamin Schmid, Jean-Yves Tinevez, Daniel James White, Volker Hartenstein, Kevin Eliceiri, Pavel Tomancak and Albert Cardona (2012). ‘Fiji: an open-source platform for biological-image analysis’. en. In: *Nat. Methods* 9.7, pp. 676–682.
- Sevillano, Víctor, Katherine Holt and José L Aznarte (2020). ‘Precise automatic classification of 46 different pollen types with convolutional neural networks’. en. In: *PLoS One* 15.6, e0229751.
- Shilatifard, Ali (2006). ‘Chromatin modifications by methylation and ubiquitination: implications in the regulation of gene expression’. en. In: *Annu. Rev. Biochem.* 75, pp. 243–269.
- Simpson, Jared T, Rachael E Workman, P C Zuzarte, Matei David, L J Dursi and Winston Timp (2017). ‘Detecting DNA cytosine methylation using nanopore sequencing’. en. In: *Nat. Methods* 14.4, pp. 407–410.
- Singh, Uday Pratap, Siddharth Singh Chouhan, Sukirty Jain and Sanjeev Jain (2019). ‘Multilayer Convolution Neural Network for the Classification of Mango Leaves Infected by Anthracnose Disease’. In: *IEEE Access* 7, pp. 43721–43729.

- Stam, Maike, Christiane Belele, Jane E Dorweiler and Vicki L Chandler (2002). ‘Differential chromatin structure within a tandem array 100 kb upstream of the maize b1 locus is associated with paramutation’. en. In: *Genes Dev.* 16.15, pp. 1906–1918.
- Sterner, D E and S L Berger (2000). ‘Acetylation of histones and transcription-related factors’. en. In: *Microbiol. Mol. Biol. Rev.* 64.2, pp. 435–459.
- Stewart, Ethan L, Tyr Wiesner-Hanks, Nicholas Kaczmar, Chad DeChant, Harvey Wu, Hod Lipson, Rebecca J Nelson and Michael A Gore (2019). ‘Quantitative Phenotyping of Northern Leaf Blight in UAV Images Using Deep Learning’. en. In: *Remote Sensing* 11.19, p. 2209.
- Stokes, Jonathan M, Kevin Yang, Kyle Swanson, Wengong Jin, Andres Cubillos-Ruiz, Nina M Donghia, Craig R MacNair, Shawn French, Lindsey A Carfrae, Zohar Bloom-Ackermann, Victoria M Tran, Anush Chiappino-Pepe, Ahmed H Badran, Ian W Andrews, Emma J Chory, George M Church, Eric D Brown, Tommi S Jaakkola, Regina Barzilay and James J Collins (2020). ‘A Deep Learning Approach to Antibiotic Discovery’. en. In: *Cell* 180.4, 688–702.e13.
- Strahl, B D and C D Allis (2000). ‘The language of covalent histone modifications’. en. In: *Nature* 403.6765, pp. 41–45.
- Sullivan, K F, M Hechenberger and K Masri (1994). ‘Human CENP-A contains a histone H3 related histone fold domain that is required for targeting to the centromere’. en. In: *J. Cell Biol.* 127.3, pp. 581–592.
- Sun, Deqiang, Yuanxin Xi, Benjamin Rodriguez, Hyun Jung Park, Pan Tong, Mira Meong, Margaret A Goodell and Wei Li (2014). ‘MOABS: model based analysis of bisulfite sequencing data’. en. In: *Genome Biol.* 15.2, R38.
- Szegedy, Christian, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke and Andrew Rabinovich (2014). ‘Going Deeper with Convolutions’. In: arXiv: 1409.4842 [cs.CV].
- Takuno, Shohei, Jin-Hua Ran and Brandon S Gaut (2016). ‘Evolutionary patterns of genic DNA methylation vary across land plants’. en. In: *Nat Plants* 2, p. 15222.
- Thomas, J O and R D Kornberg (1975). ‘An octamer of histones in chromatin and free in solution’. en. In: *Proc. Natl. Acad. Sci. U. S. A.* 72.7, pp. 2626–2630.
- Tibshirani, Robert (1996). ‘Regression shrinkage and selection via the lasso’. en. In: *J. R. Stat. Soc.* 58.1, pp. 267–288.
- Ubbens, Jordan, Mikolaj Cieslak, Przemyslaw Prusinkiewicz and Ian Stavness (2018). ‘The use of plant models in deep learning: an application to leaf counting in rosette plants’. en. In: *Plant Methods* 14, p. 6.
- Venkatesh, Swaminathan and Jerry L Workman (2015). ‘Histone exchange, chromatin structure and the regulation of transcription’. en. In: *Nat. Rev. Mol. Cell Biol.* 16.3, pp. 178–189.
- Waddington, C H (1942). ‘The Epigenotype’. In: *Endeavour* 1, pp. 18–20.
- Wang, Tao, Mina Rostamza, Zhihang Song, Liangju Wang, G McNickle, Anjali S Iyer-Pascuzzi, Zhengjun Qiu and Jian Jin (2019). ‘SegRoot: A high throughput segmentation method for root image analysis’. In: *Comput. Electron. Agric.* 162, pp. 845–854.
- Wang, Yan, Haizhen Long, Juan Yu, Liping Dong, Michel Wassef, Baowen Zhuo, Xia Li, Jicheng Zhao, Min Wang, Cuifang Liu, Zengqi Wen, Luyuan Chang, Ping Chen, Qian-Fei Wang, Xueqing Xu, Raphael Margueron and Guohong Li (2018). ‘Histone variants H2A.Z and H3.3 coordinately regulate PRC2-dependent H3K27me3 deposition and gene expression regulation in mES cells’. en. In: *BMC Biol.* 16.1, p. 107.

- Ward, Daniel, Peyman Moghadam and Nicolas Hudson (2018). ‘Deep Leaf Segmentation Using Synthetic Data’. In: arXiv: 1807.10931 [cs.CV].
- Warman, Cedar, Christopher M Sullivan, Justin Preece, Michaela E Buchanan, Zuzana Vejlupekova, Pankaj Jaiswal and John E Fowler (2021). ‘A cost-effective maize ear phenotyping platform enables rapid categorization and quantification of kernels’. en. In: *Plant J.* 106.2, pp. 566–579.
- Washburn, Jacob D, Maria Katherine Mejia-Guerra, Guillaume Ramstein, Karl A Kremling, Ravi Valluru, Edward S Buckler and Hai Wang (2019). ‘Evolutionarily informed deep learning methods for predicting relative transcript abundance from DNA sequence’. en. In: *Proc. Natl. Acad. Sci. U. S. A.* 116.12, pp. 5542–5549.
- Wassenegger, M, S Heimes, L Riedel and H L Sanger (1994). ‘RNA-directed de novo methylation of genomic sequences in plants’. en. In: *Cell* 76.3, pp. 567–576.
- Watson, J D and F H Crick (1953). ‘Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid’. en. In: *Nature* 171.4356, pp. 737–738.
- Werbos, Paul J (1982). ‘Applications of advances in nonlinear sensitivity analysis’. In: *System Modeling and Optimization*. Springer Berlin Heidelberg, pp. 762–770.
- Wick, Ryan R, Louise M Judd and Kathryn E Holt (2019). ‘Performance of neural network basecalling tools for Oxford Nanopore sequencing’. en. In: *Genome Biol.* 20.1, p. 129.
- Widman, Nicolas, Suhua Feng, Steven E Jacobsen and Matteo Pellegrini (2014). ‘Epigenetic differences between shoots and roots in Arabidopsis reveals tissue-specific regulation’. en. In: *Epigenetics* 9.2, pp. 236–242.
- Woo, Hye Ryun, Travis A Dittmer and Eric J Richards (2008). ‘Three SRA-domain methylcytosine-binding proteins cooperate to maintain global CpG methylation and epigenetic silencing in Arabidopsis’. en. In: *PLoS Genet.* 4.8, e1000156.
- Xu, Jihua, Karen K Tanino and Stephen J Robinson (2016). ‘Stable Epigenetic Variants Selected from an Induced Hypomethylated *Fragaria vesca* Population’. en. In: *Front. Plant Sci.* 7, p. 1768.
- Yang, Jianyi, Ivan Anishchenko, Hahnbeom Park, Zhenling Peng, Sergey Ovchinnikov and David Baker (2020). ‘Improved protein structure prediction using predicted interresidue orientations’. en. In: *Proc. Natl. Acad. Sci. U. S. A.* 117.3, pp. 1496–1503.
- Zhang, Xiaoyu, Junshi Yazaki, Ambika Sundaresan, Shawn Cokus, Simon W-L Chan, Huaming Chen, Ian R Henderson, Paul Shinn, Matteo Pellegrini, Steve E Jacobsen and Joseph R Ecker (2006). ‘Genome-wide high-resolution mapping and functional analysis of DNA methylation in arabidopsis’. en. In: *Cell* 126.6, pp. 1189–1201.
- Zhang, Y and D Reinberg (2001). ‘Transcription regulation by histone methylation: interplay between different covalent modifications of the core histone tails’. en. In: *Genes Dev.* 15.18, pp. 2343–2360.
- Zhang, Yuan-Ye, Markus Fischer, Vincent Colot and Oliver Bossdorf (2013). ‘Epigenetic variation creates potential for evolution of plant phenotypic plasticity’. en. In: *New Phytol.* 197.1, pp. 314–322.

List of Acronyms

ANN artificial neural network

CMT2 CHROMOMETHYLASE 2

CMT3 CHROMOMETHYLASE 3

CNN convolutional neural network

CPU central processing unit

DCL DICER-LIKE

DMR differentially methylated region

DNA deoxyribonucleic acid

DRM2 DOMAINS REARRANGED METHYLTRANSFERASE 2

EWAS epigenome wide association studies

FNN feed-forward neural network

GPU graphics processing unit

GSNO S-Nitrosoglutathione

GSNOR S-Nitrosoglutathione reductase

GUI graphical user interface

GWAS genome wide association studies

HAT histone acetyl transferase

HDAC histone deacetylase

KYP KRYPTONITE

LiDAR light detection and ranging

MET1 DNA METHYLTRANSFERASE 1

MLP multilayer perceptron

MRs methylated regions

NGS next-generation sequencing

NO nitric oxide

PCR polymerase chain reaction

RdDM RNA-directed DNA methylation

RNA ribonucleic acid

RNN recurrent neural network

SAM S-Adenosylmethionine

SNP single nucleotide polymorphism

SUVH4/5/6 SUPPRESSOR OF VARIEGATION 3-9 HOMOLOG

TE transposable element

TGS long-read sequencing

VIM1-3 VARIANT IN METHYLATION 1-3

WGBS whole genome bisulfite sequencing

XOR exclusive-or