

Advancing the analysis of bisulfite sequencing data in its application to ecological plant epigenetics

Adam Nunn



UNIVERSITÄT
LEIPZIG

Fakultät für Mathematik und Informatik

Advancing the analysis of bisulfite sequencing data in its application to ecological plant epigenetics

Von der Fakultät für Mathematik und Informatik
der Universität Leipzig
angenommene

DISSERTATION

zur Erlangung des akademischen Grades

DOCTOR RERUM NATURALIUM
(Dr. rer. nat.)

im Fachgebiet

INFORMATIK

vorgelegt

von M.Sc. Bioinf. Adam Nunn,
geboren am 20. August 1990 in Portsmouth, UK

Die Annahme der Dissertation wurde empfohlen von:

1. Professor Dr. Peter F. Stadler, Universität Leipzig, Deutschland
2. Professor Dr. Robert J. Schmitz, University of Georgia, USA

Die Verleihung des akademischen Grades erfolgt mit Bestehen
der Verteidigung am 05. September 2022 mit dem Gesamtprädikat *summa cum laude*

Contents

ABSTRACT	5
ACKNOWLEDGEMENTS	7
1 INTRODUCTION	9
1.1 ABOUT THIS WORK.....	9
1.2 BIOLOGICAL BACKGROUND	10
1.2.1 <i>Epigenetics in plant ecology</i>	10
1.2.2 <i>DNA methylation</i>	12
1.2.3 <i>Maintenance of 5mC patterns in plants</i>	14
1.2.4 <i>Distribution of 5mC patterns in plants</i>	17
1.3 TECHNICAL BACKGROUND.....	18
1.3.1 <i>DNA sequencing</i>	18
1.3.2 <i>The case for a high-quality genome assembly</i>	24
1.3.3 <i>Sequence alignment for NGS</i>	28
1.3.4 <i>Variant calling approaches</i>	32
2 BUILDING A SUITABLE REFERENCE GENOME	37
2.1 INTRODUCTION.....	37
2.2 MATERIALS AND METHODS.....	39
2.2.1 <i>Seeds for the reference genome development</i>	39
2.2.2 <i>Sample collection, library preparation, and DNA sequencing</i>	39
2.2.3 <i>Contig assembly and initial scaffolding</i>	41
2.2.4 <i>Re-scaffolding</i>	43
2.2.5 <i>Comparative genomics</i>	45
2.3 RESULTS	45
2.3.1 <i>An improved reference genome sequence</i>	45
2.3.2 <i>Comparative genomics</i>	47
2.4 DISCUSSION.....	48
3 FEATURE ANNOTATION FOR EPIGENOMICS	51
3.1 INTRODUCTION.....	51
3.2 MATERIALS AND METHODS.....	51
3.2.1 <i>Tissue preparation for RNA sequencing</i>	51
3.2.2 <i>RNA extraction and sequencing</i>	52
3.2.3 <i>Transcriptome assembly</i>	52
3.2.4 <i>Genome annotation</i>	53
3.2.5 <i>Transposable element annotations</i>	54
3.2.6 <i>Small RNA annotations</i>	54
3.2.7 <i>Expression atlas</i>	56
3.2.8 <i>DNA methylation</i>	56
3.3 RESULTS	57
3.3.1 <i>Transcriptome assembly</i>	57
3.3.2 <i>Protein-coding genes</i>	57

3.3.3 <i>Non-coding loci</i>	59
3.3.4 <i>Transposable elements</i>	60
3.3.5 <i>Small RNA</i>	60
3.3.6 <i>Pseudogenes</i>	62
3.3.7 <i>Gene expression atlas</i>	62
3.3.8 <i>DNA Methylation</i>	64
3.4 DISCUSSION.....	65
4 BISULFITE SEQUENCING METHODS.....	68
4.1 INTRODUCTION.....	68
4.2 PRINCIPLES OF BISULFITE SEQUENCING	70
4.3 EXPERIMENTAL DESIGN	74
4.4 LIBRARY PREPARATION.....	76
4.4.1 <i>Whole Genome Bisulfite Sequencing (WGBS)</i>	76
4.4.2 <i>Reduced Representation Bisulfite Sequencing (RRBS)</i>	81
4.4.3 <i>Target capture bisulfite sequencing</i>	81
4.5 BIOINFORMATIC ANALYSIS OF BISULFITE DATA.....	82
4.5.1 <i>Quality Control</i>	82
4.5.2 <i>Read Alignment</i>	84
4.5.3 <i>Methylation Calling</i>	88
4.6 ALTERNATIVE METHODS.....	89
5 FROM READ ALIGNMENT TO DNA METHYLATION ANALYSIS.....	90
5.1 INTRODUCTION.....	90
5.2 MATERIALS AND METHODS.....	92
5.2.1 <i>Reference species</i>	92
5.2.2 <i>Natural accessions</i>	92
5.2.3 <i>Read simulation</i>	93
5.2.4 <i>Read alignment</i>	94
5.2.5 <i>Mapping rates</i>	95
5.2.6 <i>Precision-recall</i>	95
5.2.7 <i>Coverage deviation</i>	96
5.2.8 <i>DNA methylation analysis</i>	96
5.3 RESULTS	97
5.4 DISCUSSION.....	106
5.5 A PIPELINE FOR WGBS ANALYSIS	108
6 THERE AND BACK AGAIN: INFERRING GENOMIC INFORMATION.....	109
6.1 INTRODUCTION.....	109
6.1.1 <i>Implementing a new approach</i>	111
6.2 MATERIALS AND METHODS.....	113
6.2.1 <i>Validation datasets</i>	113
6.2.2 <i>Read processing and alignment</i>	113
6.2.3 <i>Variant calling</i>	113
6.2.4 <i>Benchmarking</i>	114
6.3 RESULTS	114
6.4 DISCUSSION.....	121
6.5 A PIPELINE FOR SNP VARIANT ANALYSIS.....	123

7 POPULATION-LEVEL EPIGENOMICS	124
7.1 INTRODUCTION.....	124
7.2 CHALLENGES IN POPULATION-LEVEL EPIGENOMICS	125
7.3 DIFFERENTIAL METHYLATION.....	126
7.3.1 <i>A pipeline for case/control DMRs</i>	128
7.3.2 <i>A pipeline for population-level DMRs</i>	129
7.4 EPIGENOME-WIDE ASSOCIATION STUDIES (EWAS).....	130
7.4.1 <i>A pipeline for EWAS analysis</i>	132
7.5 GENOTYPING-BY-SEQUENCING (EPIGBS).....	135
7.5.1 <i>Extending the epiGBS pipeline</i>	135
7.6 POPULATION-LEVEL HAPLOTYPES.....	137
7.6.1 <i>Extending the EpiDiverse/SNP pipeline</i>	139
8 CONCLUSION.....	141
APPENDICES.....	145
A. SUPPLEMENT: BUILDING A SUITABLE REFERENCE GENOME.....	145
B. SUPPLEMENT: FEATURE ANNOTATION FOR EPIGENOMICS.....	155
C. SUPPLEMENT: FROM READ ALIGNMENT TO DNA METHYLATION ANALYSIS	164
D. SUPPLEMENT: INFERRING GENOMIC INFORMATION	165
BIBLIOGRAPHY	168

Abstract

The aim of this thesis is to bridge the gap between the state-of-the-art bioinformatic tools and resources, currently at the forefront of epigenetic analysis, and their emerging applications to non-model species in the context of plant ecology. New, high-resolution research tools are presented; first in a specific sense, by providing new genomic resources for a selected non-model plant species, and also in a broader sense, by developing new software pipelines to streamline the analysis of bisulfite sequencing data, in a manner which is applicable to a wide range of non-model plant species. The selected species is the annual field pennycress, *Thlaspi arvense*, which belongs in the same lineage of the Brassicaceae as the closely-related model species, *Arabidopsis thaliana*, and yet does not benefit from such extensive genomic resources. It is one of three key species in a Europe-wide initiative to understand how epigenetic mechanisms contribute to natural variation, stress responses and long-term adaptation of plants.

To this end, this thesis provides a high-quality, chromosome-level assembly for *T. arvense*, alongside a rich complement of feature annotations of particular relevance to the study of epigenetics. The genome assembly encompasses a hybrid approach, involving both PacBio continuous long reads and circular consensus sequences, alongside Hi-C sequencing, PCR-free Illumina sequencing and genetic maps. The result is a significant improvement in contiguity over the existing draft state from earlier studies.

Much of the basis for building an understanding of epigenetic mechanisms in non-model species centres around the study of DNA methylation, and in particular the analysis of bisulfite sequencing data to bring methylation patterns into nucleotide-level resolution. In order to maintain a broad level of comparison between *T. arvense* and the other selected species under the same initiative, a suite of software pipelines which include mapping, the quantification of methylation values, differential methylation between groups, and epigenome-wide association studies, have also been developed. Furthermore, presented herein is a novel algorithm which can facilitate accurate variant calling from bisulfite sequencing data using conventional approaches, such as FreeBayes or Genome Analysis ToolKit (GATK), which until now was feasible only with specifically-adapted software. This enables researchers to obtain high-quality genetic variants, often essential for contextualising the results of epigenetic experiments, without the need for additional sequencing

libraries alongside. Each of these aspects are thoroughly benchmarked, integrated to a robust workflow management system, and adhere to the principles of FAIR (Findability, Accessibility, Interoperability and Reusability). Finally, further consideration is given to the unique difficulties presented by population-scale data, and a number of concepts and ideas are explored in order to improve the feasibility of such analyses.

In summary, this thesis introduces new high-resolution tools to facilitate the analysis of epigenetic mechanisms, specifically relating to DNA methylation, in non-model plant data. In addition, thorough benchmarking standards are applied, showcasing the range of technical considerations which are of principal importance when developing new pipelines and tools for the analysis of bisulfite sequencing data. The complete “Epidiverse Toolkit” is available at <https://github.com/EpiDiverse> and will continue to be updated and improved in the future.

Acknowledgements

First, I would like to thank my supervisors David Langenberger and Peter F. Stadler, for their relentless support and understanding, and for their generosity in lending me both their expert advice and extensive experience when working with such a challenging research topic.

Special thanks also to Christian Otto and Mario Fasold, who were always there to provide high-quality scientific input, advice, support, and with whom I frequently enjoyed bouncing new ideas throughout this work. Thank you indeed to all my colleagues at ecSeq Bioinformatics GmbH, including Gero Doose and Adele Feuerstein, who have each helped make my time at the company both successful and enjoyable.

To everyone involved in EpiDiverse: thank you so much for your expertise, support, and kindness, both in terms of our shared experience in pursuit of scientific excellence, but also in your gracious tendency towards building-up those around you and giving them everything they need to succeed. In particular I would like to thank the ESRs: Bhumika Dubay, Paloma Pérez-Bello, Samar Fatma, Nilay Can, Iris Sammarco, Dario Galanti, Bárbara Díez Rodríguez, Morgane Van Antro, Anupoma Niloya Troyee, Maria Estefania Lopez, Daniela Ramos Cruz, Cristian Peña, Adrián Contreras Garrido and Panpan Zhang.

To so many others who were involved, including in particular my external supervisor Claude Becker, Ratan Chopra, Pablo Carbonell-Bejerano, Patrick Hüther, Fleur Gawehns, Maarten Postuma, and everyone at the University of Leipzig, including Christiane Gaertner, Jens Steuck, Petra Pregel, Sven Findeiss- thank you all!

Special thanks of course to my family and close friends, who always support me in my crazy adventures, even if it means I'm gallivanting off to another country for an undetermined number of years!

Finally, to my fiancée Stefanie, who was there at every step- this work is as much yours as it is mine, for I could not have managed it without you by my side. You always believed in me, and I will never forget it.

This thesis is based on the following publications:

Nunn A, Rodríguez-Arévalo I, Tandukar Z, Frels K, Contreras-Garrido A, Carbonell-Bejerano P, Zhang P, Ramos-Cruz D, Jandrasits K, Lanz C, Brusa A, Mirouze M, Dorn K, Jarvis B, Sedbrook J, Wyse DL, Otto C, Langenberger D, Weigel D, Marks MD, Anderson JA, Becker C, Chopra R (2022).

Chromosome-level *Thlaspi arvense* genome provides new tools for translational research and for a newly domesticated cash cover crop of the cooler climates. *Plant Biotechnology Journal*. Jan; 20:(5), p.944-963, doi:10.1111/pbi.13775

Nunn A (2021).

Bisulfite sequencing methods. In C. Lampei, K. Heer & L. Opgenoorth (Eds.), *Introduction to Ecological Plant Epigenetics*. https://epidiverse.gitbook.io/project/-MfxkdBDZggX_vc_sG5l

Nunn A, Otto C, Stadler PF, Langenberger D (2021).

Comprehensive benchmarking of software for mapping whole genome bisulfite data: from read alignment to DNA methylation analysis. *Briefings in Bioinformatics*. Feb; bbab021, doi:10.1093/bib/bbab021

Nunn A, Otto C, Fasold M, Stadler PF, Langenberger D (2022).

Manipulating base quality scores enables variant calling from bisulfite sequencing alignments using conventional Bayesian approaches. *BMC Genomics*. June; 23, p.477, doi:10.1186/s12864-022-08691-6

Can SN, Nunn A, Galanti D, Langenberger D, Becker C, Volmer K, Heer K, Opgenoorth L, Fernandez-Pozo N, Rensing SA (2021).

The EpiDiverse Plant Epigenome-Wide Association Studies (EWAS) Pipeline. *Epigenomes*. May; 5:(2) p.12, doi:10.3390/epigenomes5020012

Hüther P, Hagmann J, Nunn A, Kakoulidou I, Pisupati R, Langenberger D, Weigel D, Johannes F, Schultheiss SJ, Becker C (2022).

MethylScore: a pipeline for accurate and context-aware identification of differentially methylated regions from population-scale plant WGBS data. *Quantitative Plant Biology* [In review]. doi:10.1101/2022.01.06.475031

Gawehns F, Postuma M, van Antro M, Nunn A, Sepers B, Fatma S, van Gurp TP, Wagemaker NCAM, Mateman C, Milanovic-Ivanovic S, Grosse I, van Oers K, Vergeer P, Verhoeven KJF (2022).

epiGBS2: Improvements and evaluation of highly multiplexed, epiGBS-based reduced representation bisulfite sequencing. *Molecular Ecology Resources*. Feb, 22:(5), p.2087-2104, doi:10.1111/1755-0998.13597

Nunn A, Can SN, Otto C, Fasold M, Díez Rodríguez B, Fernandez-Pozo N, Rensing SA, Stadler PF, Langenberger D (2021).

EpiDiverse Toolkit: a pipeline suite for the analysis of ecological plant epigenetics. *NAR Genomics and Bioinformatics*. Dec; lqab106, doi:10.1093/nargab/lqab106

1 Introduction

1.1 About this work

The expanding scope and scale of next generation sequencing (NGS) experiments in the study of ecological plant epigenetics brings new challenges for computational analysis. Model organisms such as *Arabidopsis thaliana* have helped lay the foundation for much of our current understanding in regard to specific molecular pathways, mechanisms and functional consequences (Bossdorf et al. 2010; Boyko et al. 2010; Cokus et al. 2008). Now, the increasingly competitive costs of NGS have opened the door for plant ecologists to apply these lessons and gain more specific insight into non-model species, particularly on the population and community level. A recent perspective by Richards et al. (2017) highlights the need in this topic to better integrate the fields of molecular genetics and evolutionary ecology, by adding more ecological context and ecological questions to model species research (e.g. Latzel et al. 2013; Hagemann et al. 2015), and by adopting higher resolution tools in non-model species research (e.g. Platt et al. 2015; Xie et al. 2015; Gugger et al. 2016; van Gurp et al. 2016; Trucchi et al. 2016). Under the purview of the EpiDiverse project (<https://epidiverse.eu>), this thesis attempts to address the latter point; first in a specific sense, by providing new genomic resources for a selected non-model plant species, and also in a broader sense, by developing new pipelines and software tools to facilitate high-resolution analysis applicable to a wide range of non-model plant species.

The EpiDiverse consortium has identified three plant species by which to help expand the current understanding of plant epigenetics in the context of ecology and evolution. These include the deciduous broadleaf tree species black poplar, *Populus nigra* cv. 'Italica', and two flowering herbaceous plants: the perennial wild strawberry, *Fragaria vesca*, and the annual field pennycress, *Thlaspi arvense*. Each species represents a variation in life cycle, mode of reproduction, and zygosity, and are widely distributed in the wild throughout Europe. Studies are underway to investigate natural epigenetic variation, stress-response and transgenerational inheritance, and interactions among genomic elements such as transposons and populations of small RNA. Each species however has a varying level of quality in terms of available genomic resources, ranging from a chromosome-level assembly to a highly-fragmented draft, or no available genome at all. As such,

1 Introduction

the first part of this thesis encompasses efforts to improve the existing draft genome of *T. arvense*, making use of state-of-the-art technology in a hybrid assembly approach.

In addition, existing software tools built for model species may not address the needs of researchers looking to apply similar techniques to non-model species, particularly on a population or community level. The latter part of this thesis thus presents a toolkit suitable for plant ecologists working with whole genome bisulfite sequencing (WGBS) in order to study patterns in DNA methylation; it includes pipelines for mapping, the calling of methylation values and differential methylation between groups, epigenome-wide association studies, and a novel implementation for variant calling. The tools presented herein are implemented with the workflow management system Nextflow (Di Tommaso et al. 2017), building on best-practice concepts outlined by nf-core (Ewels et al. 2020). They are intended to be efficient, intuitive for novice users, optimisable for laptops, high performance computing clusters, or the cloud, and scalable from small lab studies to field trials with large populations. On a POSIX compatible system, setup is as simple as installing Nextflow itself, alongside one of either Bioconda (Grüning et al. 2018), Docker (Merkel 2014), or Singularity (Kurtzer, Sochat, and Bauer 2017), which allow for automated management of pipeline dependencies through the use of software containers and environments, thus facilitating a high level of reproducibility. The platform will be maintained and expanded upon as new tools are developed in the future.

1.2 Biological background

1.2.1 Epigenetics in plant ecology

The term “epigenetics” refers broadly to the molecular mechanisms and processes which can regulate phenotypic changes in an organism without alteration of the underlying genomic sequence. More precisely, an epigenetic mark should exhibit some level of heritability through mitotic or meiotic cell division, and be reversible (Riggs and Porter 1996). Such changes are usually driven instead by conformational differences in DNA structure and organisation, for example by histone modification, or on the nucleotide-level through base modifications such as with DNA methylation. The resulting influence on DNA-binding molecules and nuclear architecture can in turn have functional consequences for the regulation of gene expression and imprinting, cellular differentiation during development, genome stability, for example through the repression of transposable elements (TEs), and processes of nuclear maintenance such as DNA replication and repair (Richards et al. 2017; Pikaard and Mittelsten Scheid 2014; Feng, Jacobsen, and Reik 2010).

Epigenetic patterns can be either transient or stable, and much consideration has been given to the transgenerational heritability of certain epigenetic marks in reference to the evolutionary consequences under modern evolutionary synthesis. Heritable genetic variation alone is often not sufficient to explain the range of phenotypic diversity that can be observed for individuals of the same species, or within individuals that reproduce clonally, for example. This is especially true for quantitative traits, and has been referred to as the genotype-to-phenotype gap (Fernie and Gutierrez-Marcos 2019) and the missing heritability problem (Manolio et al. 2009). Epigenetic variation, albeit often confounded by genetic variation, has been recognised as another potential source of natural diversity (Riddle and Richards 2002; Cervera, Ruiz-García, and Martínez-Zapater 2002; Vaughn et al. 2007).

In contrast to traditional genetic inheritance, the concept of epigenetic inheritance has two important distinctions under its current understanding: i) the rate of stochastic epimutation (Johannes and Schmitz 2019), which can be much faster than genetic mutation, and ii) that epimutation is more easily reversible (Becker et al. 2011). Typically, this manner of heritability is related to the mode of reproduction, as in the case of DNA methylation patterns for example, many of which are preserved under cell division by mitosis but reset in the germline due to “epigenetic reprogramming” (Feng, Jacobsen, and Reik 2010). Particularly in mammals, where DNA methylation patterns are erased first during gametogenesis and again during early embryogenesis, the capacity for transgenerational epigenetic inheritance is limited (Heard and Martienssen 2014). Under sexual reproduction in flowering plants, however, meiocytes are differentiated from somatic cells, giving rise to an alternative pathway for epigenetic reprogramming whereby the methylation states of some genomic elements are perhaps more able to persist across generations (Kawashima and Berger 2014). Clonal propagation provides another means by which some plants can reproduce without germline passage at all. For example, in *A. thaliana*, a sexually reproducing plant, the epigenetic reprogramming during germline formation can be avoided through asexual reproduction, and a new epigenetic state maintained in the progeny (Wibowo et al. 2018). A broad exploration of such transitory and/or heritable effects resulting from epigenetic mechanisms across a range of plant species offers a unique insight into our understanding of evolutionary processes.

In the context of plant ecology, epigenetic research is more specifically concerned with i) patterns of natural epigenetic variation, ii) the origins and drivers of this variation, and iii) its ecological and

1 Introduction

evolutionary consequences (Richards et al. 2017). These questions centre chiefly around the study of “epigenomics” in natural populations, the interplay between genetic and epigenetic variation, and the influence (if any) of the surrounding environment. Current insight, however, is built upon a foundation of understanding from the molecular genetics of model organisms, such as *A. thaliana*. It remains unclear whether previous experimental findings hold true given the added complexity of natural conditions, particularly for non-model species, and it is a challenge to adapt existing methods to the expanded scope and scale of evolutionary ecology. This thesis seeks to address this challenge, first by generating new genomic resources for non-model species, and second by the implementation of novel software pipelines, with particular emphasis on the analysis of sequencing data intended to study patterns of DNA methylation.

1.2.2 DNA methylation

DNA methylation is most often characterised by the covalent attachment of a methyl group from S-adenosylmethionine to the carbon 5 of a cytosine nucleotide, thus converting it to 5-methylcytosine (5mC) (Pikaard and Mittelsten Scheid 2014; Sahu et al. 2013). This epigenetic mechanism is itself involved in a broad range of molecular processes such as gene regulation (Jaenisch and Bird 2003), transposon silencing (Miura et al. 2001), and heterochromatin formation (Lippman et al. 2004); it plays a role in genome organisation (Zemach et al. 2013; Huff and Zilberman 2014), developmental processes (Finnegan, Peacock, and Dennis 1996; Ronemus et al. 1996; Papareddy et al. 2021) and imprinting (Li et al. 1993; Kinoshita et al. 2004; Köhler et al. 2005; Gehring et al. 2006; Jullien et al. 2008; Pignatta et al. 2018). DNA methylation can also be considered a source of phenotypic variation, following the observation of natural variation among populations (Eichten et al. 2011; Heyn et al. 2013; Schmitz et al. 2013).

In many eukaryotic organisms, cytosine (C) methylation typically occurs in a “CG” sequence context, i.e. when followed by a guanine (G). This form of DNA methylation is the most predominant in nature, and yet there is considerable variation in its genomic organisation between taxonomic kingdoms. In mammals, unmethylated CG dinucleotides tend to be concentrated together in so-called “CpG islands” (where “p” denotes the phosphodiester bond joining the two nucleotides), which are non-randomly distributed along the genome and play an important role in transcriptional regulation (Deaton and Bird 2011). The dinucleotides populating the remaining, “CG-deficient” fraction of the genome on the other hand are globally methylated. Plant genomes however typically exhibit “mosaicism”, i.e. regions of interspersed methylated and unmethylated domains, which can vary substantially between species (Suzuki and Bird 2008). Plant DNA

methylation is also further complicated by its prevalence in two additional sequence contexts: CHG and CHH (where H is any base but G). Methylation in each context is governed by an independent molecular machinery, dedicated to establishing, maintaining, and regulating it (Henderson and Jacobsen 2007; Law and Jacobsen 2010). The frequency of methylation along the genome therefore differs accordingly by context; in the model plant *A. thaliana*, for example, CGs are methylated most frequently (~24% of all CGs), followed by CHG (~6.7%) and CHH (~1.7%) relative to the total number of cytosines in each context (Cokus et al. 2008).

The study of whole-genome methylation profiles across the plant and animal kingdoms has revealed both conserved and divergent features of DNA methylation in eukaryotes (Feng et al. 2010; Zemach et al. 2010). For example, CG methylation within protein-coding genes is preferentially concentrated in exons, which appears to be a conserved feature predating the divergence of plants and animals. In flowering plants (*Arabidopsis*, rice, and poplar), all three contexts (CG, CHG, and CHH) also show similar patterns, with high enrichment in repetitive DNA, transposons, and pericentromeric regions (Feng et al. 2010). DNA methylation is nevertheless highly variable in genome-wide levels and distribution, both within and between species, even down to the level of individual tissues and developmental stages. This variation has consequences for a variety of molecular processes. Some phenotypic traits, for example, including floral symmetry (Cubas, Vincent, and Coen 1999), fruit development and morphology (Manning et al. 2006; Zhong et al. 2013; Ong-Abdullah et al. 2015), plant height (Miura et al. 2009), and resistance to pathogens (Liégard et al. 2019), have been associated with naturally occurring epigenetic alleles (epialleles) in absence of linked genetic variants. Dysfunctions of stable DNA methylation patterns in plants have also been shown to lead to abnormalities, such as fruit ripening failure in tomato and orange, or vice versa to promote early fruit ripening in strawberry (Cheng et al. 2018; H. Huang et al. 2019; Lang et al. 2017).

While genomic origin and sequence context broadly underpin the stability of 5mC patterns on a genetic basis in accordance with specific molecular mechanisms, there is yet further evidence from both plants and animals that patterns of DNA methylation variation can be inherited independently through mitosis and, at least partially, also through meiosis (Chong and Whitelaw 2004; Richards 2006; Henderson and Jacobsen 2007; Law and Jacobsen 2010). Indeed, this transgenerational inheritance has been demonstrated for example in *A. thaliana* through the use of “epigenetic Recombinant Inbred Lines” (epiRILs), unveiling DNA methylation as a possible source of heritable phenotypic variation whereby epialleles can influence complex traits in the absence of

1 Introduction

DNA sequence change (Johannes et al. 2009; Reinders et al. 2009). Moreover, mutation accumulation lines in *A. thaliana* have revealed that in addition to stable plant DNA methylation patterns which are able to persist over many generations, the rate of stochastic epimutation is both higher than can be explained by the (lower) rate of spontaneous genetic mutation, and far more susceptible to reverse epimutation (Becker et al. 2011; Schmitz et al. 2011). Understanding how epialleles become triggered and/or released under this apparent “transgenerational instability” will provide insight as to the evolutionary consequences arising from variation in DNA methylation, for example as a possible mechanism for local adaptation.

Furthermore, DNA methylation variation is apparently responsive to environmental stimuli (Jaenisch and Bird 2003; Lloyd and Lister 2022). One notable example occurs during vernalisation in *A. thaliana*, the process by which flowering can be triggered in response to prolonged cold (winter) temperatures. The transcription factor FLOWERING LOCUS C (FLC) is a repressor of genes important for the transition to flowering, and its gradual silencing is the driving mechanism for inducing this change (Sheldon et al. 2000). Interestingly, a reduction in DNA methylation has been shown to regulate FLC in this manner (Sheldon et al. 1999), raising the postulation that phenotypic plasticity can be mediated in a “controlled” manner on an epigenetic basis. In the years that followed, the situation was discovered to be markedly more complex however, with vernalisation seeming to occur independently of DNA methylation (Jean Finnegan et al. 2005) but nevertheless still under epigenetic control (Bastow et al. 2004; De Lucia et al. 2008). Even so, there remain several less well-studied examples of the environment modulating plant development through DNA methylation (Lloyd and Lister 2022). Changes in DNA methylation brought about by biotic or abiotic stresses are of particular interest, with regards to whether such environmentally-induced cues can be passed on through the germline to later generations (Secco et al. 2015; Wibowo et al. 2016). A complete understanding of the extent by which such a mechanism for local adaptation might occur in nature has thus far proven elusive, however (Alonso, Ramos-Cruz, and Becker 2019).

1.2.3 Maintenance of 5mC patterns in plants

Global DNA methylation patterns are responsible in part for regulating genome stability through the silencing of transposons, preserving cell type identity during development, and even to establish an “epigenetic memory” against environmental stresses, for example (Law and Jacobsen 2010). It is therefore essential for plants to maintain them. Typically, the regulation of genome-wide DNA methylation patterns is carried out by DNA methyltransferases, which differ by sequence context

(CG, CHG, and CHH) and may form cooperative or competing interactions (Meyer 2011; Zhang, Lang, and Zhu 2018). The efficiency of these different DNA methyltransferases is reflected in the methylation level at their preferred target sites. At the single position level, based on the proportion of overlapping sequencing reads which contribute to the estimation, methylated cytosines in CG context are frequently found to be 80-100% methylated, for example, whereas those methylated in CHG context vary uniformly between 20-80%, and those in CHH context peak at 10-30% (Cokus et al. 2008; Lister et al. 2008).

Even beyond the described dinucleotide and trinucleotide contexts, the more specific sequence context can yet further influence DNA methylation efficiency in plants. For example, CG sites are undermethylated when the exact four-base context is ACGT (Cokus et al. 2008; Lister et al. 2008). In non-CG context, the cytosine is less efficiently methylated when followed by another cytosine, as in CCG context for example (Gouil and Baulcombe 2016). These more specific contexts seem to modulate the efficiency of DNA methyltransferases, though the reason is not well understood. For example, in *A. thaliana*, different methyltransferases share similar sequence specificities to perhaps provide a methylation backup system and avoid harmful alterations in the plant genome (Law and Jacobsen 2010; Li et al. 2018; Meyer 2011).

DNA methylation in CG and CHG context is symmetrical, i.e. the complementary sequences on the opposite strand mirror the methylation state. CHH methylation, on the other hand, is asymmetric and occurs on just one strand at each particular locus. The consequence of this is reflected mainly in the capacity to re-establish methylation on the daughter strand following semi-conservative DNA replication, whereby DNA methyltransferases operating on symmetric sequence contexts can be recruited depending on the methylation state of the original strand (Law and Jacobsen 2010; Zhang, Lang, and Zhu 2018).

Maintenance of CG methylation. In plants, CG methylation is maintained via the methyltransferase METHYLTRANSFERASE 1 (MET1) that, akin to its mammalian homolog DNA METHYLTRANSFERASE 1 (DNMT1), shows substrate preference to hemimethylated DNA (Kankel et al. 2003; Saze, Mittelsten Scheid, and Paszkowski 2003; Jeltsch 2006). To do this, MET1 is probably recruited to the replication complex by the VARIANT IN METHYLATION (VIM) protein family of SRA (SET- and RING-associated) domain proteins (Law and Jacobsen 2010; Zhang, Lang, and Zhu 2018), following the observation that VIM1 loss-of-function mutants lose the DNA methylation of their centromeres in *A. thaliana* (Kim et al. 2014). The chromatin-

1 Introduction

remodelling protein DECREASED DNA METHYLATION 1 (DDM1) is also important for maintaining methylation in CG context, wherein loss of DDM1 function has been shown to cause a 70% reduction of genomic cytosine methylation (Jeddeloh, Stokes, and Richards 1999), though it is involved in the methylation of other contexts as well.

Maintenance of non-CG methylation. In contrast to CG methylation, non-CG methylation is maintained by plant-specific enzymes in the CHROMOMETHYLASE (CMT) family (Kenchanmane Raju, Ritter, and Niederhuth 2019), and depends on additional factors such as histone modifications and small RNAs. The DNA methyltransferases CMT3 and, to a lesser extent, CMT2 catalyse the maintenance of CHG methylation in *A. thaliana* (Zhang, Lang, and Zhu 2018), and are reciprocally dependent on H3K9 dimethylation (Jackson et al. 2002; J. Du et al. 2012). Interestingly, CMT3 appears to have a role also in gene-body methylation (gbM) in several plant species. This has been demonstrated in *Eutrema salsugineum*, a plant that naturally lacks both gbM and CMT3, where the experimental gain of CMT3 triggered a new establishment of gbM in genes homologous to naturally-methylated genes in *A. thaliana*, which were maintained even in following generations (Wendte et al. 2019).

Asymmetric DNA methylation. Without opposite-strand information by which to re-establish DNA methylation patterns in CHH context, maintenance is solely dependent on mechanisms for *de novo* methylation instead. The methyltransferases CMT2 (Zemach et al. 2013; Stroud et al. 2014) and DOMAINS REARRANGED METHYLTRANSFERASE 2 (DRM2), which also facilitate *de novo* methylation in symmetric sequence contexts, are as of yet the only known mechanisms responsible for CHH methylation. DRM2 is mediated by small RNAs through the RNA-directed DNA methylation (RdDM) pathway (Aufsatz et al. 2002; Cao and Jacobsen 2002; Cao et al. 2003; Wassenegger et al. 1994; Matzke and Mosher 2014), whereby it is recruited to RdDM target regions, preferentially located at transposons or other repeat sequences in the euchromatin, as well as the edges of long transposons usually located in the heterochromatin (Zhang, Lang, and Zhu 2018). CMT2 on the other hand targets histone H1-containing heterochromatin sites, where RdDM is inhibited, in coordination with DDM1 (Zemach et al. 2013; Zhang, Lang, and Zhu 2018). The varying extent of specificity by which each of these mechanisms are involved in the methylation of different sequence contexts demonstrates an underlying natural complexity, whereby each sequence context cannot truly be considered truly independent of one another.

1.2.4 Distribution of 5mC patterns in plants

Given that different molecular mechanisms are responsible for methylation of different sequence contexts, the distribution of methylation patterns along the genome therefore is expected to differ accordingly by context. (Cokus et al. 2008) demonstrated in the model plant *A. thaliana*, for example, that CGs are methylated most frequently (~24% of all CGs), followed by CHG (~6.7%) and CHH (~1.7%) relative to the total number of cytosines in each context. Moreover, it is important to note that - at least in some plant species, including *A. thaliana* - DNA methylation differs also in terms of the methylation level between the different sequence contexts, i.e. in the consistency of the methylation status of a given cytosine across different cells. CG cytosines for instance tend to have an almost binary methylation state, being either unmethylated (0%) or methylated (i.e 80-100%) in almost all cells of a given tissue (Cokus et al. 2008; Lister et al. 2008). 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.

Furthermore, the distribution of 5mC patterns is species-specific and often dependent on neighbouring genomic features. Plant genomes are highly dynamic, with tremendous variation in content, size and complexity, brought about by multiple evolutionary processes including whole-genome duplication events, the proliferation and loss of lineage-specific transposable elements, and various classes of small RNAs which help shape genomic architecture and function (Wendel et al. 2016). This in turn corresponds with the diversification of DNA methylation patterns between species. Plants typically adhere to a mosaic pattern of 5mC distribution characterised by regions of interspersed methylated and unmethylated domains (Suzuki and Bird 2008), whereby the DNA is often methylated at the body of genes (where the function is often unclear) and at repetitive regions, where it restricts the expression of TEs, which represent in some plant species more than 80% of the genome, e.g. barley, sunflower, and maize (Meyer 2011; Vitte et al. 2014). Indeed, those with larger genomes due to extensive proliferation of TEs can appear almost akin to species from other kingdoms which are globally methylated.

Among flowering plants, genome-wide methylome studies (Niederhuth et al. 2016) have highlighted how differences in aspects such as gene body DNA methylation, euchromatic silencing of transposons and repeats, as well as the silencing of heterochromatic transposons, can be reflective of evolutionary and life histories among clades. (Niederhuth et al. 2016) demonstrate for example that the Brassicaceae have generally reduced CHG methylation levels, and also reduced or lost CG gene body methylation, whereas the Poaceae are characterised by a lack or reduction of

heterochromatic CHH methylation and enrichment of CHH methylation in genic regions. Furthermore, low levels of CHH methylation were observed in a number of species, especially in those which undergo clonal propagation. The variation in 5mC patterns between different plant species opens new areas of study to understand the role of DNA methylation and its correlation with evolutionary distance as well as biological diversity (Niederhuth et al. 2016; Feng et al. 2010; Zemach et al. 2010).

1.3 Technical background

1.3.1 DNA sequencing

In order to garner insight into the mechanisms linking both genetics and epigenetics to changes in phenotype, it is necessary to study DNA at the nucleotide-level. Sequencing technology provides the means by which to observe variations in DNA at such a resolution, and has rapidly evolved over the years since its conception into a number of techniques with a wide range of applications (Metzker 2010). One of the first approaches, the Sanger sequencing method (Sanger, Nicklen, and Coulson 1977; Sanger et al. 1977), involves the use of both standard deoxynucleotides (dNTPs) and dideoxy-modified dNTPs (ddNTPs). Following a standard PCR reaction, any random incorporation of a ddNTP during synthesis of a new DNA strand results in its termination, inhibiting further elongation. The resulting product therefore contains a number of newly-synthesised DNA fragments at a range of different sizes, terminated at various stages of synthesis, which can then be separated accordingly by gel electrophoresis. Given four concurrent PCR reactions, each with a homogenous mixture of dNTPs but with only one type of ddNTP (e.g. ddATP, ddCTP, ddGTP, or ddTTP), the resulting fraction from each nucleotide can be run side-by-side on the polyacrylamide gel in order to infer the original template sequence by comparison. This strategy was a pioneer of “sequencing-by-synthesis”, and is still adopted even today due to its exceedingly low error rate (~99.99% accuracy). Modern variations often make use of radioactive or fluorescent labels to differentiate ddNTPs within a single PCR reaction (Smith et al. 1986; Prober et al. 1987), but the method overall is yet limited to a low yield of short sequences up to approximately 100-1000 bp.

1.3.1.1 Next-generation sequencing

The next major development in sequencing technology was brought about by the arrival of next-generation sequencing (NGS). In contrast to early sequencing approaches, significant advances in overall yield were made at the expense of sequenced fragment length and the accuracy of individual

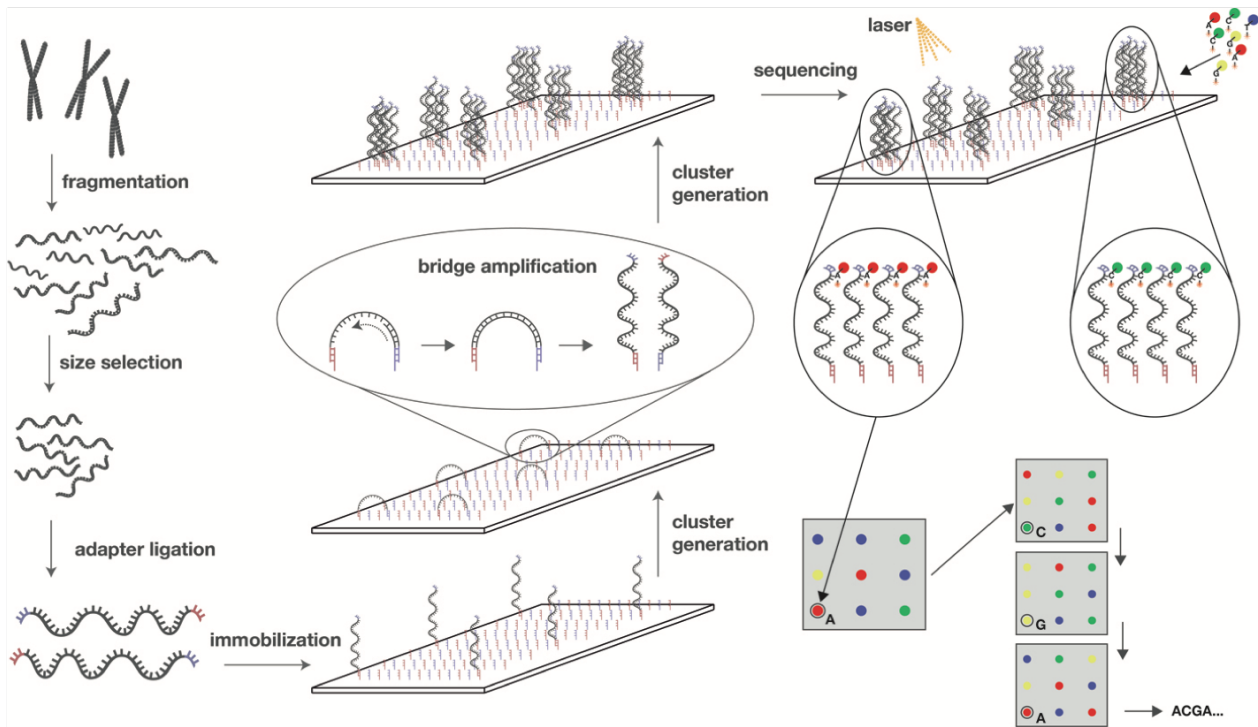


Figure 1. Under Illumina sequencing, DNA is fragmented, size selected, then immobilized to a flow cell by a process of adapter ligation. Bridge amplification by PCR enables the generation of clusters, which are then sequenced-by-synthesis following a cyclic reversible termination method. At each incorporation of a new nucleotide, a resulting fluorescent signal is captured by the imaging system in order to perform base calling.

base calls (Metzker 2010; Goodwin, McPherson, and McCombie 2016). Perhaps the most prominent among these NGS approaches is the cyclic reversible termination (CRT) method of Illumina (Figure 1), now the market leader of NGS technologies in comparison to other platforms. Similar to the Sanger method in that it is sequencing by synthesis, the CRT method affixes each DNA template to a “flow cell” by adapter ligation, forming clusters of identical sequences which are then elongated in a cyclical manner through a process of “reversible termination”, using dNTPs which are blocked at the ribose 3'-OH group to prevent the next dNTP from binding. The unbound dNTPs are then washed from the flow cell, and the incorporated dNTP in each cluster can be identified by excitation of a bound fluorophore molecule, emitting a specific wavelength for each nucleotide which can be captured by an imaging system. The fluorophore and blocking group can then be removed and a new cycle can begin. As of a review by (Goodwin, McPherson, and McCombie 2016), the throughput on Illumina machines ranges from ~0.5-900 Gb, producing individual “sequencing reads” which are equal to the number of cycles (~25-300 bp) - a vast improvement on overall yield in comparison to Sanger sequencing. Yet more modern machines such as the NovaSeq 6000 are even capable of a throughput of up to ~6000 Gb with 150 bp paired-end reads (Levy and Boone 2019). The error profile for Illumina sequencing ranges from a rate of

1 Introduction

<0.1% to <1% incorrect base calls, which is among the best for NGS methods (Goodwin, McPherson, and McCombie 2016).

Further methods of sequencing by synthesis include single nucleotide addition (SNA), utilised for example in 454 pyrosequencing and Ion Torrent machines, which typically result in longer reads than Illumina, but lower throughput. Alternatively, “sequencing by ligation” is a method employed for example by SOLiD (Sequencing by Oligonucleotide Ligation and Detection) and Complete Genomics. These approaches depend on fluorescent-labelled probes with complementary nucleotides which are hybridised to the DNA template on a slide, and ligated to an upstream anchor sequence. Following ligation, the slide is imaged to identify the nucleotide sequence based on its associated label and the probes cleaved. This series of probe-anchor binding, ligation, imaging and cleavage is repeated in each instance following single-nucleotide offsets to ensure every base in the DNA template is sequenced. Sequencing reads from SOLiD and Complete Genomics typically have comparable throughput to Illumina, but reduced read length (Goodwin, McPherson, and McCombie 2016).

The limitations of NGS methods thus far manifest predominantly in the reliability of individual base calls, and the length of the sequencing reads which are produced. Once sequencing is completed, it becomes a non-trivial bioinformatic task to determine how the library of sequencing reads correspond to the original contiguity of the sequenced DNA. Short reads originating from highly-repetitive regions in particular are a challenge for both downstream alignment and assembly approaches, which are common next steps in sequencing analysis, due to the ambiguity arising from the numerous potential loci which a given read could have come from. This can be mitigated for example through the use of “paired-end” sequencing. Under the Illumina method, each DNA template can be ligated at both ends with different adapters, and the entire sequencing procedure repeated in each case. Given that the DNA template is affixed to the same cluster on the flow cell in both instances, this approach results in the formation of “mate pairs”, of equal length, which correspond to either end of the same DNA fragment. Taken together with the known distribution of sequenced DNA fragment lengths, this information can be used to infer a plausible distance between two mate pairs which thus reduces the level of ambiguity during downstream alignment and assembly procedures.

1.3.1.2 Single-molecule long-read sequencing

The benefit of longer sequencing reads for applications such as read alignment, RNA transcript reconstruction and assembly has given rise to third-generation sequencing approaches such as Pacific Biosciences (PacBio) (Eid et al. 2009; Uemura et al. 2010) and Oxford Nanopore Technologies (ONT) (Cherf et al. 2012; Mikheyev and Tin 2014). Unlike Illumina, for example, read length is not limited by the number of cycles for reversible termination, which rapidly accumulate in errors beyond a short read length. Instead, read length is more limited to the size of the DNA template which can be obtained by chemical means during the specific library preparation procedures carried out in the lab. PacBio is similar to Illumina in that it is a sequencing by synthesis approach, making use of specialised flow cells (known as “SMRT cells”) whereby instead of affixing the DNA template through adapter ligation, the DNA polymerase molecule which catalyses the elongation is affixed at the base of a picolitre well known as a “zero-mode waveguide” (ZMW) (Figure 2). The template itself is ligated at each end with single-stranded hairpin adapter molecules, forming a “SMRTbell template” which is subsequently circular in nature. Together with their associated sequencing primers, the hairpin adapters provide a starting point for the DNA polymerase to begin elongation of the complementary sequence on either strand. As the polymerase incorporates fluorophore-labelled dNTPs to the template, a fluorescence pulse is produced by the polymerase retaining the cognate nucleotide with its colour-coded fluorophore in

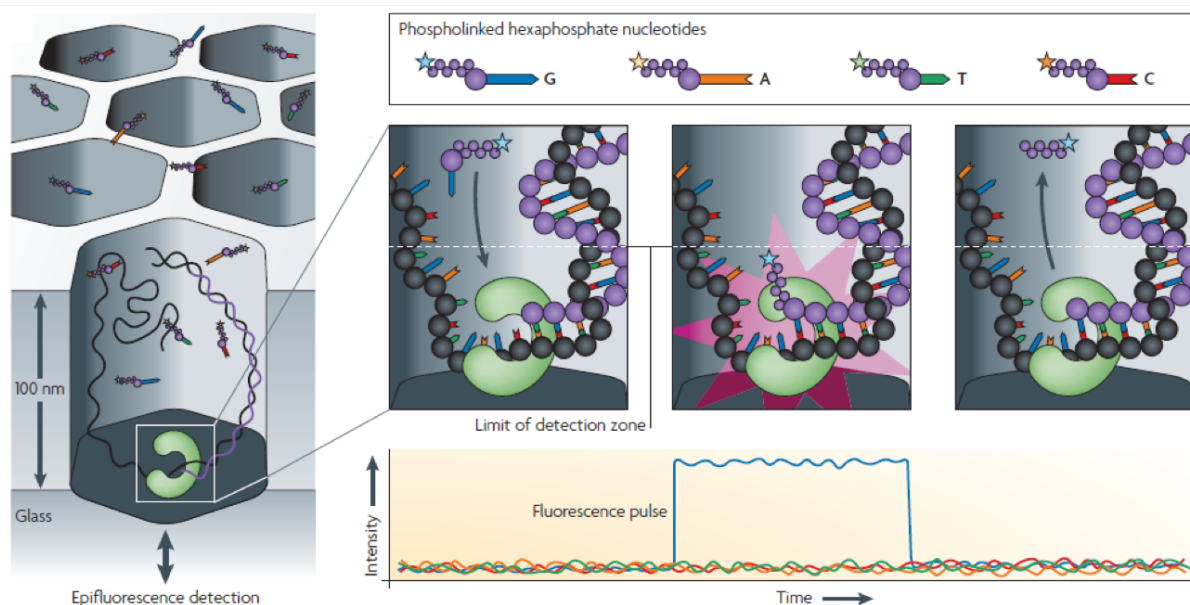


Figure 2. Under PacBio sequencing the DNA polymerase is transfixed to the base of each picolitre well, known as a zero-mode waveguide (ZMW), on the flow cell. Base calling is performed in real time as each nucleotide incorporated to the sequencing releases a fluorophore which is captured by the imaging system. Adapted from Metzker (2010) “Sequencing technologies – the next generation”.

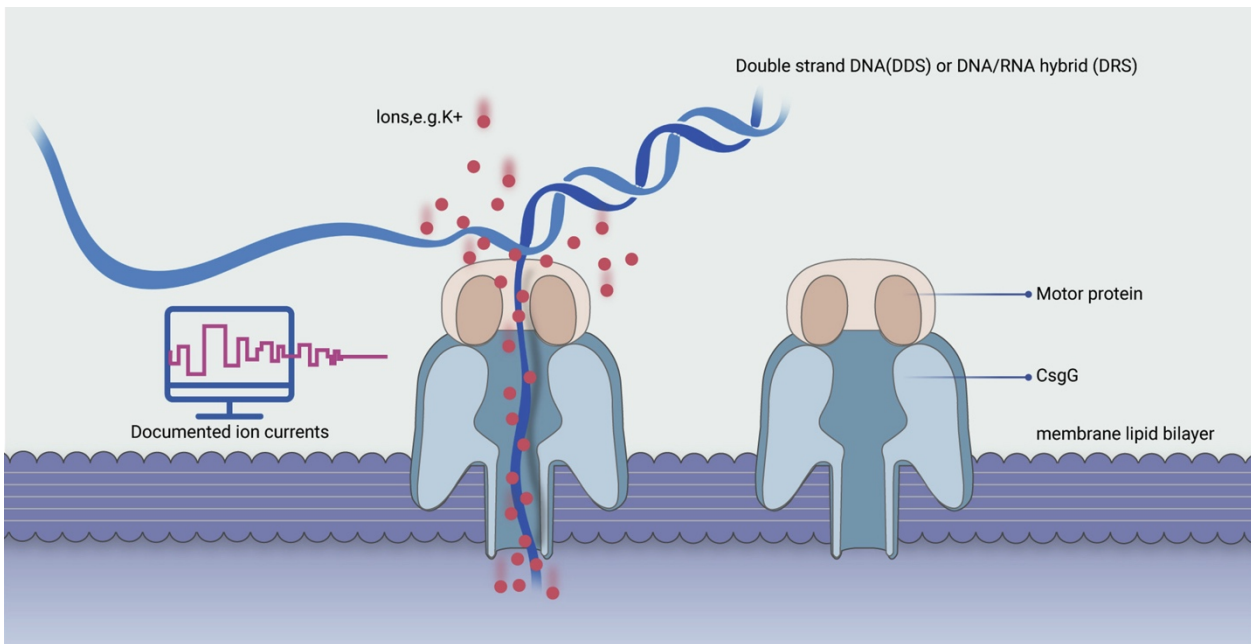


Figure 3. Under Oxford Nanopore sequencing, an electric current differential is applied across a membrane lipid bilayer. A motor protein guides the DNA sequencing through a nanopore, which modulates the signal measured from the electric current according to the k-mer subsequence present at any given time in the nanopore. Base calling is derived from the documented signal known as the “squiggle space”. Adapted from Xie et al. (2021) “Applications and potentials of nanopore sequencing in the (epi)genome and (epi)transcriptome era”.

the detection region of the ZMW, which ends upon cleavage of the dye-linker-pyrophosphate group. The colour and duration of emitted light is captured by an imaging system in real time. Interestingly, the kinetic variation emitted from the light signal during the base calling can even reveal base modifications such as DNA methylation (Flusberg et al. 2010). Typically, multiple passes of the same circular SMRTbell template are made to form an overall consensus of the underlying sequence, known as “circular consensus sequences” (CCS), which reduces the error rate in single base calls to <1%, down from approximately 10-15% in comparison to sequencing a single pass from one hairpin adapter to another, i.e. a “continuous long read” (CLR). High-fidelity CCS reads are thus more accurate, but tend to reliably sequence shorter insert sizes of 10-20 kbp, whereas CLR reads can sequence longer reads typically in the range of 25-175 kbp. The current Sequel II technology has a throughput of ~160 Gb per SMRT cell.

ONT sequencing on the other hand employs a distinct approach which does not follow either sequencing by synthesis or by ligation (Figure 3). Nanopore technology instead takes a single stranded DNA molecule and passes it through a protein pore, wherein an electric current is applied and modulated according to the native composition of the DNA as it translocates from nucleotide to nucleotide through the action of a secondary motor protein. The signal trace (known as the

“squiggle space”) emitted in real time is characteristic to the particular k-mer DNA sequence (3-6 bp) that is measured from the pore at that moment, according to the shifts in voltage that arise. One of the main difficulties is in reliably interpreting the emitted trace at single nucleotide resolution, but conversely this characterisation is in principle not only limited to the four nucleotides and can be extended even to identifying base modifications, similar to PacBio (Xie et al. 2021). As base calling algorithms improve in the future, this could have implications for the study of DNA methylation and epigenetics for example.

1.3.1.3 FASTQ format and the concept of base quality

In order to represent sequencing reads for computational processing, it is necessary to define a file format which is applicable for reads arising from any sequencing technology. The most common format for this purpose is FASTQ, whereby each read is represented by four lines: 1) a sequence identifier, beginning with an “@” character, often containing other metadata pertaining to the sequencing run, 2) the literal sequence itself (e.g. “ACTGTG...”), 3) some information regarding the base qualities, and 4) a string of ASCII characters of equal length to the sequence, representing each individual base quality score associated with each nucleotide.

Base qualities scores are given as a measure of confidence for whether the base caller involved in the sequencing experiment made the correct call during the sequencing run, based on the information provided to it. As per equation (1), it is a phred-scaled probability estimate Q defined as a property which is logarithmically related to the base-calling error probability P (Ewing and Green 1998).

$$Q = -10 \log_{10}(P) \quad \text{equation (1)}$$

Scores are seen as positive integers typically in the range of 0 - 40 where higher values reflect greater confidence in the base call. Common values are assigned to single ASCII characters to aid in compression and computational interoperability, whereby they can then be interpreted during downstream processes such as sequence alignment and variant calling. It is important to note that the score pertains only to errors which occur during sequencing and are not necessarily reflective of real errors in the sequence itself. In other words, a nucleotide with a high phred-score is indicative only that the base caller had no other conflicting signal by which to interpret it. Errors for example which occur prior to sequencing however would not be captured, and thus cannot be conveyed by this measure.

1.3.2 The case for a high-quality genome assembly

When presented with a large number of short sequences from high-throughput sequencing, the next task is to determine how they correspond to their genomic point of origin. If there is no existing reference genome available to facilitate this, then it is often necessary to first assemble sequences back together in an order which is representative of the original unfragmented DNA. Only by establishing the correct original sequence can we make accurate biological inferences in regard to genome arrangement and architecture, and thus genome assembly is a pivotal step upon which almost all downstream analysis is dependent. Low quality draft or incomplete assemblies have a tendency to contain errors, frequently due to limitations in sequencing technology or in assembly algorithms. Such errors typically manifest as base errors, collapsed or expanded repeat regions, or rearrangements and inversion (Phillippy, Schatz, and Pop 2008), which can then impact later analysis for example in terms of estimated gene content (Florea et al. 2011; Denton et al. 2014), split genes and misorientations (Xiongfei Zhang, Goodsell, and Norgren 2012), and missing genome content resulting in overall reduced genome size (Alkan, Sajjadian, and Eichler 2011). The quality of the reference genome is thus directly correlated to the quality of its feature annotations and any subsequent downstream analysis, and thereby represents a natural first step in any study seeking to make inferences from such information (e.g. when studying DNA methylation). Low quality, draft or even absent reference genome sequences are all the more likely in non-model species.

There are many different approaches to assembly (Wee et al. 2019; Li et al. 2011; Li and Harkess 2018; Schatz, Witkowski, and McCombie 2012; Bradnam et al. 2013; Rhie et al. 2021), which differ in suitability depending for example on species, estimated genome size, and the NGS technology used to generate the read library. From beginning to end, the process is usually multi-faceted, typically involving not one single procedure but a workflow of iterative steps such as read correction, contig assembly, scaffolding, polishing, gap filling, and phasing. The necessity of each step is determined on a case-by-case basis according to project specification, resources, and also on the outcome of each previous step; ergo it is unusual for any genome assembly initiative to be comparatively optimal under different use cases. All approaches begin from the same principle, however, in that NGS reads beyond a certain level of sequencing coverage can be linked together based on shared, overlapping subsequences.

1.3.2.1 *De novo* assembly approaches

Genome assembly can be generalised as either i) reference-based, or ii) *de novo*. Reference-based approaches make use of existing genome sequences, either of the same or closely-related species, in order to help reconstruct the most likely order of sequences from NGS reads. This is typically more appropriate for pan-genome studies, where for example the aim is to investigate structural variants among different accessions of the same species and a high-quality reference genome is already available. This approach is often less prone to error than *de novo* methods, but errors that do occur can be compounded due to the dependence on the quality of what was assembled before. Conversely, *de novo* methods attempt to assemble the genome from scratch.

Contig assembly. *De novo* approaches typically employ a variation of either overlap-layout-consensus (OLC) or de Bruijn graphs (DBG) to assemble short sequences into longer “contigs” (Li et al. 2011). OLC methods begin first by generating all possible pairwise alignments of reads, to form an “overlap graph”, i.e. a directed multigraph whereby each read represents a node and each resulting overlap represents an edge, weighted according to length. The overlap graph represents the layout, and the sequence is determined by finding the Hamiltonian path through the graph which minimises the overall weight. This last step is computed iteratively to form a number of plausible sequences, and a consensus achieved for example using multiple sequence alignment. In contrast, DBG methods work instead on the principle of using k-mers, whereby each sequencing read is first reduced to its overlapping component subsequences of length k. Every k-mer in the read is then further subdivided into two component subsequences of length k-1, and every unique k-1-mer in the dataset is then represented by a single node in the graph. Each k-mer from each read represents an edge in the graph, linking its two-component k-1-mer nodes. The overall sequence is found by traversing through the graph following the Eulerian path, i.e. wherein each edge is visited exactly once. Both OLC and DBG methods struggle to a varying extent when attempting to resolve repetitive sequences, particularly when such regions are longer than the corresponding sequencing reads used to assemble them. The increased complexity of long-read sequencing is therefore one way to improve genome assembly, leading to fewer, longer, more reliable contig sequences and thus a greater “contiguity” of resulting assemblies.

Scaffolding. After a series of contiguous sequences have been assembled, often there is still a degree of fragmentation between them which makes it difficult to make inferences regarding the overall genomic landscape. Whereas a “complete” genome assembly would contain a set of sequences equal to haploid chromosome number, frequently there are yet hundreds or even

1 Introduction

thousands of independent sequences obtained from contig assembly. Several approaches exist so as to reorient and order these contigs into something resembling the overall genome, including for example i) optical and/or genetic linkage maps, ii) chromosome conformation capture (3C), and iii) synteny-based comparison to related species. All work on the basis of identifying genetic markers which convey some level of distance or spatial organisation between them, which can then be used to form a representative arrangement of contigs. For example, Hi-C sequencing is a form of chromosome conformation capture, whereby first “DNA cross-links” are formed by chemical treatment of formaldehyde in the lab, representing interacting genomic loci with proximity in 3D space. The DNA is then fragmented by enzymatic digestion, and subject to proximity-based ligation which thus favours cross-linked DNA. Ligated DNA cross-links are subsequently enriched, the cross-link itself removed, and the resulting DNA subject to paired-end sequencing. In this scenario, each mate pair is expected to originate from the opposing genomic loci on either side of the initial DNA cross-link interaction. Software tools such as SALSA (Ghurye et al. 2017) make use of mate pairs alignments to assembled contigs in order to build up a genome-wide interaction map which reflects their spatial organisation, thus facilitating the arrangement of contigs into scaffolds.

Polishing. Often when assembling contigs using long-read sequencing technology, a trade-off is made in terms of the individual base quality of constituent nucleotides. Current long-read sequencing technology (e.g. PacBio, ONT) is comparatively more error-prone than short-read sequencing with Illumina, for example. Hybrid approaches to genome assembly therefore seek to utilise different libraries together to leverage advantages from both, for example by aligning short-read sequences to contigs assembled with long-read technology, and correcting nucleotides by consensus overlap. Software such as Pilon (Walker et al. 2014) have been developed to handle short-read libraries in this manner, whereas others such as Racon (Vaser et al. 2017) exist to facilitate polishing with long reads such as PacBio CCS for example.

1.3.2.2 Feature annotation

Once a suitable reference genome is assembled and available, the next task is to make sense of the complete sequence in context of the underlying biology. Genes, repeat sequences, transposable elements, RNA loci and pseudogenes are all examples of features which may need to be annotated in the reference genome in order to provide a frame of reference for further downstream analysis. There are many different software tools for genome annotation, which differ primarily by which feature they intend to annotate, and further still by whether their objective is to identify structure

or exact function. Most approaches operate on a basis of sequence homology, i.e. given a set of known sequences, what inferences can be made regarding similarity or relatedness to corresponding sequences in the genome?

Gene prediction. The first priority for every genome assembly project is to identify gene structures in the resulting sequence. In eukaryotic genomes such as those of plants, genes are encoded in an intron/exon structure whereby each exon represents a protein-coding sequence enclosed by a start codon and a stop codon. Ab initio gene predictors can utilise this information in order to carry out a *de novo* annotation of “gene models”, in principle without the need of an existing set of sequences in order to make comparisons. In practice, however, gene model parameters differ by species and are difficult to estimate, necessitating in most cases the use of machine-learning algorithms and training datasets. One such example is GeneMark-ES (Lomsadze et al. 2005), which uses an unsupervised, iterative Viterbi training algorithm to estimate hidden Markov model (HMM) parameters directly from the genome sequence itself. This was among the first methods to apply an unsupervised approach to eukaryotic genomes, resulting in similar or even greater accuracy in comparison to supervised methods on model species such as *Arabidopsis thaliana*, *Caenorhabditis elegans* and *Drosophila melanogaster*. Given that an appropriate set of experimentally-validated gene sequences, expressed sequence tags (EST) of related species, and/or sequenced cDNA libraries are available, however, it is typically more robust to include them in the estimation of gene model parameters. Methods which rely only on sequence homology are nevertheless limited by the availability of existing data at an appropriate evolutionary distance. Modern ab initio predictors such as AUGUSTUS (Stanke and Waack 2003; Stanke et al. 2006) instead combine an accurate, unsupervised approach, capable of resolving splice variants, with the option for increased specificity using e.g. ESTs or cDNA libraries either as “hints” or by directly training the algorithm.

1.3.2.3 FASTA, BED and GTF/GFF formats

The most common format to denote DNA sequences computationally is FASTA, whereby each sequence (e.g. gene, contigs, scaffolds, or chromosomes) is represented by 1) a single line pertaining to a sequence identifier, beginning with a “>” character, preceding 2) one or more lines which encompass the literal sequence itself (e.g. “ACTGTG...”). In contrast to the FASTQ format, these are the only lines necessary to define within the file. It may often be accompanied by a FASTA index file, representing a dictionary wherein each line refers to a sequence ID from the corresponding FASTA file and its associated length.

1 Introduction

When referring to specific loci associated with a sequence in a FASTA file, often BED (browser extensible data) format is used. BED format refers to a tab-delimited text file wherein each locus is denoted by a single line, beginning with a column referring to a sequence ID from the FASTA, and then two columns referring to the numerical start and end position of the region, respectively. The coordinates are 0-based i.e. the first nucleotide in a sequence would carry a start position of 0 and end position of 1. The first three columns are mandatory, but further columns (name, score, etc.) are yet defined for additional detail.

Gene transfer format (GTF) and general feature format (GFF3) are commonly used to represent nested feature annotations such as genes. Similar to BED files, they are tab-delimited text files which make reference to specific loci in a corresponding FASTA file. Each line refers to a specific feature. The first column refers to the sequence ID, followed by one column denoting the “source” of the described feature and one denoting the “type”. The start and end coordinates are given in columns 4 and 5, and are 1-based i.e. the first nucleotide in a sequence would carry a start position of 1 and an end position of 1. The main difference between GTF and GFF is in the format variation of the 9th column, wherein each feature is associated with an ID which allows them to be nested with other features. For example, a feature line describing a coding sequence (CDS) may be associated with the parent ID of the gene it belongs to.

1.3.3 Sequence alignment for NGS

When a reference genome assembly is made available for a given study species, it can be used as a basis for estimating the point of origin for re-sequenced reads. There are many applications for re-sequencing; from studying relatedness and cataloguing the variation between individuals, to understanding somatic or germline variants and their effect on disease or phenotype, quantifying mRNA transcripts in regard to gene expression, identifying microorganisms within clinical or environmental samples, and in understanding the molecular mechanisms for regulating genes and the proteins they encode. In all such cases, sequence alignment plays a significant role in the inference of biologically relevant information from sequenced DNA (Metzker 2010; Li and Homer 2010).

1.3.3.1 Pairwise sequence alignment

The most basic task in sequence alignment is the comparison of two sequences to one another such that their differences are reduced to the fewest and therefore most likely divergent events. It

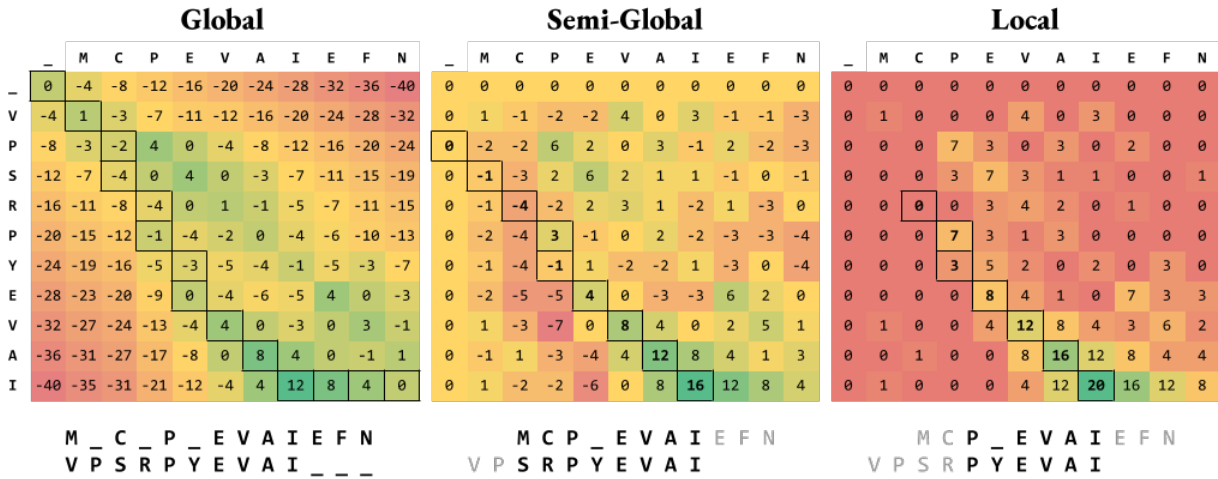


Figure 4. Examples of “Global”, “Semi-Global”, and “Local” pairwise alignment strategies on the same sequences, derived in each case from a backtracing procedure on an iteratively-generated scoring matrix.

is often these differences which present the most relevance in understanding the underlying biology, such as those for example which arise through mutation, insertion or deletion, which are by nature rare events and thus drive the principle that the more closely-related two sequences are, the fewer isolated differences there should be. There are many approaches to finding the optimal alignment, but they generally conform to three basic types: global, semi-global and local (Figure 4). Global alignment refers to the approach first popularised by Needleman and Wunsch (Needleman and Wunsch 1970), which attempts to solve the problem for similarly-sized sequences by aligning each individual character. It achieves this by denoting each sequence as an axis on a “scoring matrix”, whereby matches, mismatches and indels correspond to difference scores, and each cell is evaluated recursively for the highest score based on the result of neighbouring cells. More specifically, each iteration in the matrix F is denoted by equation (2), wherein $F(i,j)$ is the entry in the i -th row and j -th column. The matrix is first initialised with $F(0,0) = 0$ for the first cell, then $F(i,0) = F(i-1,0) - d$ and $F(0,j) = F(0,j-1) - d$ where d refers to the gap penalty. The top row represents a deletion, the middle row an insertion, and the final row a match/mismatch, wherein $s(x_i,y_j)$ refers to the score for aligning the bases x and y .

$$F(i,j) = \max \begin{cases} F(i-1,j) - d \\ F(i,j-1) - d \\ F(i-1,j-1) + s(x_i,y_j) \end{cases} \quad \text{equation (2)}$$

1 Introduction

The recursion is terminated when the whole matrix is filled, upon reaching the final cell in the bottom-right corner. The optimal alignment can then be derived through a process of “backtracing” from the lattermost cell in the bottom-right corner of the grid, by traversing in direction of the next adjacent cell with the highest score. Until a complete alignment is formed, diagonal traversal indicates a match or mismatch, whereas vertical and horizontal traversal is indicative of an indel. This method is still used even today when the priority is to identify the best possible alignment between two exact sequences.

By contrast, semi-global alignment may be more appropriate for partially overlapping sequences, or when one sequence is much shorter than another. In such a case, the Needleman-Wunsch algorithm is modified to remove penalties from the scoring matrix at either end of the sequences, sometimes referred to as end-gap free alignment. In this case the matrix is instead initialised with $F(i,0) = 0$ and $F(0,j) = 0$ and terminated upon reaching either the bottom row or the right-hand column, but each iteration is otherwise exactly the same as in equation (2).

Local alignment differs still further, in that the two sequences are expected to be largely dissimilar in overall context, perhaps differing partially in size and having only localised regions or sequence motifs which are somehow comparable between them. In such a case it may be inappropriate to obtain an optimal alignment along the entire length of either sequence against one another. An extension to the Needleman-Wunsch algorithm was therefore put forward, which allows for arbitrary length insertions and deletions and to find new alignments whenever a negative score is achieved, such as to identify the optimal pair of subsequences with the greatest homology from two larger sequences (Smith and Waterman 1981). In this case, each iteration over the matrix F is instead represented by equation (3), which is initialised in the same manner as for semi-global alignment.

$$F(i, j) = \max \begin{cases} 0 \\ F(i-1, j) - d \\ F(i, j-1) - d \\ F(i-1, j-1) + s(x_i, y_j) \end{cases} \quad \text{equation (3)}$$

In this method an additional case is included to the scoring matrix, where zero is a possible score to denote no similarity between two components. The backtracing also begins instead at the cell

with the maximal score, allowing for optimal alignments which occur anywhere along the length of either sequence. Modern applications of local alignment vary from identifying regions that may indicate functional, structural and/or evolutionary relationships between two biological sequences, to the fast lookup of databases using heuristic approaches such as BLAST (Altschul et al. 1990).

1.3.3.2 Short-read sequence alignment

With the advent of NGS technologies, the task of obtaining the optimal alignment applies to a given sequencing read against an appropriate reference genome. This is then further complicated by the high throughput of sequencing and the sheer volume of reads, in which it is no longer feasible to utilise aforementioned alignment techniques on such a scale. Most software instead make use of heuristics in order to facilitate a more rapid approximation of the most likely points of origin for each read against the genome, in a process often referred to as “read mapping”. More conventional alignment techniques can then be applied on a more targeted scale, in what amounts to a trade-off which typically comes at the expense of overall precision and sensitivity. There are many short-read mapping software which have been developed for various different applications of NGS. As of 2012, an extensive survey identified 60 different tools (Fonseca et al. 2012) and today it is thought to be more than 90. In almost all cases the mapping task is achieved through the use of index structures, such as hash tables, Burrows-Wheeler transform, and suffix arrays/enhanced suffix arrays, in order to facilitate rapid traversal to genomic loci. These data structures are often referenced using a matching “seed”, i.e. a short but representative k-mer sequence which is thus used to reduce the overall search space to a manageable number of candidate loci. When a candidate is found, the algorithm attempts to extend the alignment, which can be either successful or unsuccessful based on various scoring thresholds and an estimation of alignment “quality”. Sensitivity is typically therefore co-dependent on the exactness of these seed sequence matches, which can themselves be subject to natural mutations or errors arising from sequencing, for example, and the criteria by which each software is programmed to circumvent this. In addition, short-read aligners have to contend not only with the ambiguities arising from various sources of error but also with the nature of short sequences which originate from repetitive regions. Though each read must have originated from only one point of origin, it is possible for two or more equally good alignments to different loci to exist. In terms of the “optimal” (i.e. best-scoring) alignments, the tool may therefore i) report only “unique” alignments, thus discarding multi-mapping reads, ii) choose one or more “random” best alignment(s) in cases with multiple best-scoring alignments, or iii) report “all” best-scoring alignments. The tool may also choose to report “sub-optimal” alignments which can refer either to those with equal best scores that were

not randomly selected, or “all-first-N” alignments whereby N-1 of the next best-scoring alignments in descending order are also given. Reporting is also dependent on the search space, where typically the aligner may either opt to either search “all” candidate alignments, or abandon the search after a number of consecutive attempts to extend a candidate seed which fail to result in an equal or better alignment.

1.3.3.3 Sequence Alignment/Map (SAM) format

An important consideration for all short-read sequence aligners is in the representation and interpretation of alignments for downstream analysis. For this reason, the SAM file format was devised by (Li et al. 2009) and has since been widely adopted as a universal standard for this type of data. It encompasses first a header, containing for example file-level metadata, a reference sequence dictionary to represent scaffold/contig lengths, read group data to differentiate sequencing reads, and program-level metadata for the appropriate sequence aligner. Following the header are a series of tab-delimited lines, each representing a distinct, individual alignment, with mandatory fields to describe the exact position, sequence, characteristics and the composition of the alignment itself. Of key importance is the understanding of the bitwise flag, which contains information pertaining to the read mate status and strand orientation of the alignment, alongside various quality control considerations. A series of unordered “tags” (column 12+) can also denote characteristic information regarding the alignment, for example the NM tag which informs of the “edit distance” with respect to an identical alignment, or the NH tag which informs about the total number of reported hits associated with the corresponding read. The sequence itself is always given in a left-to-right orientation in respect of the alignment to the provided reference genome, which is assumed to be the “forward” strand, as opposed to the 5'-to-3' orientation of the read expected from the respective FASTQ file. The corresponding base quality string denoting the likelihood of individual base calls is given in the same orientation as the sequence. Finally, the MAPQ field denotes the overall quality score of the alignment itself, which is estimated according to the specification of the software used to align the read, and typically represents the phred-scaled likelihood that the given alignment is correct based on the available information.

1.3.4 Variant calling approaches

One of the most common applications of re-sequencing with NGS data is to understand the biological significance of small differences between the sample and the corresponding reference genome. Such differences can present for example as single nucleotide polymorphisms (SNPs) or short insertions/deletions (indels), which often arise naturally as point mutations in the germline

before later becoming segregated in a population. SNPs represent the most common form of genetic variant throughout the genome; estimated to occur approximately every 1 in 300 bp in the human genome (1000 Genomes Project Consortium et al. 2015) and every 1 in 10 bp in *Arabidopsis thaliana* (1001 Genomes Consortium et al. 2016). Within a given population, a SNP will often have two or more common alleles which occur at different frequencies, typically denoted in context of the “minor allele frequency” (MAF), i.e. the frequency of the less common allele. Variants with a very low MAF (e.g. <1%) can be considered rare mutations, which may not yet have segregated through the population, and are thus not considered as true SNPs but often referred to instead more generally as single nucleotide variants (SNVs).

Though for simplicity SNPs are often considered to be born of “random mutations”, it is now more generally accepted that they occur somewhat stochastically under influence of nucleotide composition, epigenomic features and genomic biases in DNA repair (Monroe et al. 2022). Emerging variants can have varying consequences on fitness depending on the context of the genomic position in which they occur; as for example with “non-synonymous” variants which can cause change or loss of function in protein-coding genes, or by introducing premature stop codons (stop-gain), or by disrupting open-reading frames or splice-sites (Xu et al. 2019). Such mutations can also instead be “synonymous”, wherein the coded protein remains the same due to codon degeneracy. Variants occurring in non-coding or intergenic regions may be less likely to cause deleterious effects, but can still influence nearby gene expression through interaction with transcription factors and other genomic machinery. Typically, mutations in coding regions have a greater likelihood to be deleterious than not, and the resulting strength of the effect on fitness is correlated with natural selection. Strongly deleterious mutations will be quickly purged from a population by purifying selection, whereas weakly deleterious mutations may yet persist given the standing load capacity of the species. Conversely, advantageous mutations which are adaptive will be more likely to accumulate under selection.

In the context of ecological plant epigenetics, SNPs are useful DNA markers both in terms of i) molecular biology, whereby they aid in understanding the function of genes and proteins potentially under influence of epigenetic effects, and ii) population ecology and evolution, whereby they serve as a measure of relatedness between individuals which may or may not be correlated with patterns of epigenetic diversity. It is therefore a desirable bioinformatic task to detect sequence variants from re-sequenced NGS reads in order to build a more comprehensive picture of the genetic background. There are many software which have been developed to perform this analysis

beginning from short-read sequence alignments, including notable examples such as bcftools (Li 2011), GATK (McKenna et al. 2010), FreeBayes (Garrison and Marth 2012), Platypus (Rimmer et al. 2014), and VarScan (Koboldt et al. 2009). The basic principle is to compare each overlapping read, for each position of the genome, where there is enough sequencing coverage to make a confident estimate of the different alleles which may be present and their respective frequencies. This in turn allows for an estimate of the “genotype” for each tested individual of the population, in respect to the number of chromosome copies (i.e. ploidy) of the species.

1.3.4.1 Bayesian approaches

A very typical approach to variant calling makes use of Bayes’ probability theorem. Under a simple Bayesian framework for example, the conditional probability of observing the true genotype G given the variants observed in the sequencing data D can be represented as per equation (4), which formulates the problem as the derivation of a prior estimate of the genotype $P(G)$ and the likelihood of observing the data $P(D|G)$.

$$P(G|D) = \frac{P(G)P(D|G)}{\sum_i P(G_i)P(D|G_i)} \quad \text{equation (4)}$$

Any assumptions made in the estimation of priors, or further extensions of the equation to incorporate various different factors or coefficients, often underlie the main differences between different Bayesian approaches. For example, given that NGS data is seldom error-free, even the simplest model will typically incorporate base quality (BQ) information directly into the Bayesian inference of genotypes as a fundamental scaling factor for the data likelihood estimation. The standard BQ score itself is a phred-based quality value which denotes on each position the estimated probability that the base caller identified the correct nucleotide during sequencing. Read alignments with low BQ scores thus typically carry less weight in such a model when the overall call of the genotype is estimated.

Alignment-based approaches. The simplest case is represented by naive approaches such as bcftools and GATK UnifiedGenotyper, in the sense that each locus is considered independently. Over each position, the corresponding alignments are “piled-up” and the likelihood estimation calculated based only on the directly aligned nucleotides and their characteristics. The aforementioned tools were among the first developed for high-throughput variant calling, and

differ only marginally for example in handling of low MAPQ reads during pre-processing and in the assumption of independence for NGS errors.

Haplotype-based approaches. The assumption of independence in the estimation of genotype likelihood for each locus often does not hold, for example due to linkage disequilibrium (LD), i.e. the nonrandom association of alleles which are inherited together. More proximal variants are more likely to be associated for example due to genetic linkage and/or recombination. Haplotype-aware variant callers, such as FreeBayes, extend the alignment-based approach to perform genotype likelihood estimation instead based on observed haplotypes, i.e. the literal sequence of each read alignment is considered as opposed to each component nucleotide individually. As NGS reads are each sequenced from a single DNA fragment, all base nucleotides are expected to arise from the same haplotype and thus the local “phasing” of co-occurring alleles can be leveraged. This is more robust particularly in terms of short indels, which can be more reliably resolved in this way by reducing the influence of ambiguous alignments. The Bayesian model is otherwise comparable to alignment-based approaches in terms of prior estimates and taking into consideration errors in the form of e.g. MAPQ and BQ scores. Notably, it is applicable to both multiallelic loci and non-uniform copy number and thus able to model genotype likelihoods from non-diploid species (Garrison and Marth 2012).

Local assembly-based approaches. Modern variant callers such as GATK HaplotypeCaller have extended further upon the principle of haplotype-based variant calling, by introducing local re-assembly and re-alignment steps to better resolve longer haplotypes and improve the signal-to-noise ratio. It operates in four steps: 1) first identifying a subset of loci based on evidence of sequence variation; 2) reassembly of each identified “ActiveRegion” from the aligned reads into plausible haplotypes, followed by alignment of each haplotype to the reference genome to identify potential variant sites; 3) the re-alignment of each read to each haplotype, in order to establish first which haplotypes are supported by which reads, followed then by the per-read likelihoods of alleles on potential variant sites; 4) Bayesian inference of genotype likelihoods per sample, given the observed likelihoods of alleles from the read data (Poplin et al. 2018).

1.3.4.2 Variant Call Format (VCF)

Just as SAM format represents a standardised approach for representing NGS sequence alignments in the same way from different software, VCF has been developed in order to provide a standard file format for variant calls. Like SAM format, it begins with a header containing file-level metadata,

1 Introduction

a reference sequence dictionary to represent scaffold/contig lengths, and definitions for various attributes that are displayed for example in the “INFO” and “FORMAT” columns in the main file. There are typically at least ten columns in the main file for variants obtained from a single sample, with one more column for each additional sample that may have been included in the analysis. The non-specific columns pertain to information relevant to a given SNP in the context of the given population, including for example genomic coordinates, reference allele and possible alternative alleles. The sample columns contain more information regarding the observations specific to each sample, for example the counts of reference and each alternative allele, genotype likelihoods and the called genotype itself.

2 Building a Suitable Reference Genome

2.1 Introduction

Within the EpiDiverse initiative, three non-model species were chosen to help broaden our understanding of ecological plant epigenetics beyond the context of model species such as *Arabidopsis thaliana*. These include the deciduous broadleaf tree species black poplar (*Populus nigra* cv. 'Italica'), and two flowering herbaceous plants: the perennial wild strawberry (*Fragaria vesca*), and the annual field pennycress (*Thlaspi arvense*). Each species represents a variation in life cycle, mode of reproduction, and zygoty, and are widely distributed in the wild throughout Europe. As with many non-model species, however, the availability of high-quality genomic resources is variable. Indeed, when the species were selected at the beginning of the project only *F. vesca* was reported to have a chromosome-level genome assembly available (Edger et al. 2018). For *P. nigra* only the genome of a related species had been published (Tuskan et al. 2006). In the case of *T. arvense*, a low quality draft genome had been published which comprised 31,889 contigs arranged into 6,768 scaffolds, representing only ~64% of the estimated genome size (Dorn et al. 2015). Such a fragmented genome makes it difficult to infer meaningful relationships in terms of genome-wide association and genetic linkage, which need to be resolved before any independent observations of epigenetic variation can be verified. Improving the quality and contiguity of the *T. arvense* genome was therefore the first priority in lieu of further planned downstream analyses.

Field pennycress is a member of the Brassicaceae family, and is closely related to the oilseed crop species *Brassica rapa*, *Brassica napus*, *Camelina sativa*, as well as the wild plant *A. thaliana* (Beilstein et al. 2010; Warwick, Francis, and Susko 2002). It is a homozygous diploid species ($2n = 2x = 14$) (Mulligan 1957) and is predominantly self-pollinating (Mulligan and Kevan 1973), suggesting a possible amenability for transgenerational epigenetic inheritance, for example in terms of stress response. Indeed, recent studies have now shown the effect of salinity stress on epigenetic diversity in populations of *T. arvense*, which were partially transferred to at least two generations of offspring (Geng et al. 2020). Within EpiDiverse the main objectives are to study the influence of DNA methylation in terms of genetic and environmental drivers of large-scale epigenetic variation, stress memory and the extent (if any) of transgenerational inheritance, and to describe populations of

2 Building a Suitable Reference Genome

small RNA-interacting loci and transposable elements (TEs) which may contribute to epigenetic variation.

Aside from its relevance as a study species in EpiDiverse, the pennycress seed is also notable for an oil content (~30-35%) and fatty acid profile conducive to producing biofuels (Fan et al. 2013; Moser 2012; Moser et al. 2009) with yet further potential to be converted into an edible oil and protein source (Chopra et al. 2020; Claver et al. 2017; McGinn et al. 2019). As such, it has garnered interest as an emerging oil feedstock species with the potential to improve sustainability of cold climate cropping systems through use as a cash cover crop (Boateng, Mullen, and Goldberg 2010; Chopra et al. 2018; Sedbrook, Phippen, and Marks 2014). Pennycress is extremely winter hardy (Warwick, Francis, and Susko 2002) and can be planted in fallow periods following traditional summer annuals such as wheat, maize or soybean (Cubins et al. 2019; Johnson et al. 2015; Ott et al. 2019; Phippen and Phippen 2012). By providing a protective living cover from the harvest of the previous summer annual crop through early spring, pennycress prevents soil erosion and nutrient loss, which in turn protects surface and below-ground water sources, suppresses early-season weed growth, and provides a food source for pollinators (Eberle et al. 2015; Johnson et al. 2015; Weyers et al. 2021; Weyers et al. 2019). The short life cycle allows for harvest even in late spring in temperate regions, with reported seed yields ranging from 750 to 2400 kg ha⁻¹ (Cubins et al. 2019; Moore et al. 2020). Following harvest, an additional crop of summer annuals can be grown in a double crop system that provides increased total seed yields and beneficial ecosystem services (Johnson et al. 2015; Phippen and Phippen 2012; Thomas et al. 2017).

The diploid and self-pollinating nature of field pennycress suggest that breeding efforts could proceed with relative ease and speed. It is amenable to genetic transformation via the floral dip method (McGinn et al. 2019), and with many one-to-one gene correspondences with *A. thaliana* (Chopra et al. 2018) it could provide an avenue for gene discovery followed by field-based phenotypic validation. Indeed, several agronomic and biochemical traits have already been identified using this translational approach, including traits crucial for *de novo* domestication such as transparent testa phenotypes (Chopra et al. 2018), early flowering (Chopra et al. 2020), reduced shatter (Chopra et al. 2020), and seed oil composition traits (Chopra et al. 2020; Esfahanian et al. 2021; Jarvis et al. 2021; McGinn et al. 2019). Such an amenability for translational research constitutes a clear advantage, which is both similar to the existing model species *A. thaliana* and at close enough evolutionary distance wherein it remains somewhat feasible to leverage model resources for comparative genomics. Field pennycress could thus serve as both i) a *de*

novoo domesticated oilseed crop for the cooler climates of the world, and ii) a new dicotyledonous model for functional genetics studies and epigenetic research.

To establish *T. arvensis* as both a research model and a new crop species, it is important to develop genomic resources that will help explore the spectrum of genetic diversity, the extent and patterns of gene expression, genetic structure, and untapped genetic potential for crop improvement. This chapter describes a set of new resources developed for both research and breeding communities, including a high quality, chromosome-level genome assembly of *T. arvensis* var. MN106-Ref, representing ~97.5% of the estimated genome size of 539 Mbp, an NG50 of 64.9 Mbp and an LG75 which reflects the haploid chromosome number of seven. Comparative genomics with closely-related Brassicaceae demonstrates a high level of synteny based on a set of ancestral chromosomal blocks (ABKs) defined by (Murat et al. 2015), revealing the extent of genomic rearrangement in coding regions. Further annotations to the genome will be described in chapter 3.

2.2 Materials and methods

2.2.1 Seeds for the reference genome development

Seeds from a small natural population of *T. arvensis* accession MN106 were collected near Coates, Minnesota by Dr. Donald L. Wyse. A single plant was propagated for ten generations from this population and is herein referred to as MN106-Ref.

2.2.2 Sample collection, library preparation, and DNA sequencing

2.2.2.1 PacBio CLR library

MN106-Ref plants were cultivated, sampled and prepared at the Max Planck Institute for Developmental Biology (Tübingen, Germany). Plant seeds were stratified in the dark at 4°C for 4-6 d prior to planting on soil. Samples of young rosette leaves were collected from seedlings, cultivated for two weeks under growth chamber conditions of 16-23°C, 65% relative humidity, and a light-dark photoperiod of 16h:8h under 110-140 $\mu\text{mol m}^{-2} \text{s}^{-1}$ light. High molecular weight (HMW) DNA was obtained following nuclei isolation and DNA extraction with the Circulomics Nanobind Plant Nuclei Big DNA kit according to the protocol described in Workman et al. (Workman et al. 2019; 2018). A total of 11 extractions from 1.5-2 g frozen leaves each were processed in that way, yielding a pooled sample with a total of 12 μg of DNA by Qubit® 2.0 fluorometer (Thermo Fisher Scientific, Waltham, MA, USA) estimation, and high DNA purity with

2 Building a Suitable Reference Genome

a mean absorbance ratio of 1.81 at 260/280 nm absorbance and 2.00 at 260/230 nm absorbance, as measured by Nanodrop 2000/2000c spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA). HMW DNA was sheared by one pass through a 26G needle using a 1 mL syringe, resulting in an 85 kb peak size sample as estimated by FEMTO Pulse Analyzer (Agilent Technologies, Santa Clara, CA, USA). A large insert gDNA library for PacBio Sequel II CLR sequencing was prepared using the SMRTbell® Express Template Preparation Kit 2.0. The library was size-selected for >30 kb using BluePippin with a 0.75% agarose cassette (Sage Science) and loaded into one Sequel II SMRT cell at a 32 pM concentration. This yielded a genome-wide sequencing depth of approximately 476X over ~6.9 million polymerase reads with a subread N50 of ~38 kbp.

2.2.2.2 PacBio CCS library

MN106-Ref plants were grown in growth chambers at the University of Minnesota. Individual plants were grown to form large rosettes for isolating DNA. Approximately 25 g of tissue was harvested and submitted to Intact Genomics (Saint Louis, MO, USA) for High Molecular Weight DNA extraction. This yielded a pooled sample with a total of 269 ng of DNA by Qubit® (Thermo Fisher Scientific, Waltham, MA, USA) estimation, and high DNA purity with a mean absorbance ratio of 1.87 at 260/280 nm and 2.37 at 260/230 nm, as measured by Nanodrop spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA). A Salt:Chloroform Wash protocol recommended by PacBio was used to further clean up the high molecular weight DNA. This yielded a total of 12.1 ng/μl of high quality DNA for library preparation. A large insert gDNA library was prepared, and 15 kb High Pass Size Selection on Pippin HT was performed at the University of Minnesota Genomics Centre (Minneapolis, MN, USA). The resulting libraries were then loaded onto four SMRT cells for sequencing with PacBio Sequel II (Pacific Biosciences, Menlo Park, USA).

2.2.2.3 Hi-C library

The same MN106-Ref plant tissue used for PacBio CCS was submitted to Phase Genomics (San Diego, CA, USA). The Hi-C library was prepared following the proximo Hi-C plant protocol (Phase Genomics, San Diego, CA, USA) and the libraries were sequenced to 116X depth on an Illumina platform in paired-end mode with read length of 150 bp.

2.2.2.4 Bionano library

HMW DNA was isolated from young leaves and nicking endonuclease - BspQI was chosen to label high-quality HMW DNA molecules. The nicked DNA molecules were then stained as

previously described (Lam et al. 2012). The stained and labelled DNA samples were loaded onto the NanoChannel array (Bionano Genomics, San Diego, CA, USA) and automatically imaged by the Irys system (Bionano Genomics, San Diego, CA, USA).

2.2.2.5 Illumina PCR-free library

Libraries for PCR-free short-read sequencing were prepared from MN106-Ref genomic DNA using the TruSeq DNA PCR-Free Low Throughput Library Prep Kit (Illumina, San Diego, CA, USA) in combination with TruSeq DNA Single Indexes Set A (Illumina, San Diego, CA, USA) according to the manufacturer's protocol. Two libraries were prepared, with average insert sizes of 350 bp and 550 bp, respectively. Samples were sequenced to ~125X depth (~66 Gb) on an Illumina HiSeq2500 (Illumina, San Diego, CA, USA) instrument with 125 bp paired-end reads.

2.2.3 Contig assembly and initial scaffolding

2.2.3.1 Genome size estimation using flow cytometry and k-mer based approach

The nuclei of field pennycress line MN106-Ref, *A. thaliana*, Maize, and Tomato were stained, with propidium iodide and fluorescent signals were captured using a Becton-Dickinson FACSCanto flow cytometer (<https://www.bdbiosciences.com/>). DNA content for all four species that corresponded to G0/1 nuclei are listed in Suppl. Table A.1. The genome size of Arabidopsis is 135 Mb and therefore, the genome size of pennycress was calculated to be 501 ± 33 Mb. Using the Illumina HiSeq2500 platform, ~100X PCR-free reads were obtained which were used for subsequent k-mer analysis using Jellyfish (Marçais and Kingsford 2011). The 101-mer frequency distribution curve exhibited a peak at 22-kmer and analysis showed that the total number of k-mers was 11,403,836,319. Using the formula of genome size = total k-mer number/peak depth, the genome size of this sequencing sample was estimated to be 518,356,196 bp. Similarly, the single copy content of the genome was estimated to reach 79%. Using both methods of genome size estimation, the pennycress genome was considered to range from approximately 459 to 539 Mb.

2.2.3.2 Contig assembly and deduplication

The final procedure for contig assembly and scaffolding was achieved using an iterative workflow which was curated over time by both trial and error and the incorporation of new sources of data. This resulted in an initial assembly from PacBio CLR reads, performed using Canu v1.9 (Koren et al. 2017) with mostly default options aside from cluster runtime configuration and the settings `corOutCoverage=50`, `minReadLength=5000`, `minOverlapLength=4000`,

`correctedErrorRate=0.04` and `genomeSize=539m`, which were selected based on the observed characteristics of the library. Canu performs consensus-based read correction and trimming, resulting in a curated set of reads which were taken forward for assembly (Suppl. Figure A.1). The resulting assembly overestimated the genome size by approximately 53% (Suppl. Table A.2), which we surmised was likely due to uncorrected sequencing errors in the remaining fraction of reads which Canu was able to assemble into independent, duplicated contigs. Analysis of single-copy orthologs from the Eudicotyledons odb10 database with BUSCO v3.0.2 (Simão et al. 2015) revealed a high completeness of 98.4% and a duplication level of 23.6% (Suppl. Table A.3). Subsequent alignment of the reads to the assembly using `minimap2 v2.17` (Li 2018) and `purge_dups v1.0.1` (Guan et al. 2020) presented bimodal peaks in the read depth distribution, indicative of a large duplicated fraction within the assembly (Suppl. Figure A.2). As efforts to collapse this duplicated fraction using assembly parameters were unsuccessful, and `purge_dups` is intended to correct duplication arising from heterozygosity (which does not apply in *T. arvensis*), the fraction was reduced by manual curation instead. Contigs starting from the left-hand side of the read depth distribution were consecutively removed until reaching an approximation of the estimated genome size, with any contigs containing non-duplicated predicted BUSCO genes kept preferentially in favour of discarding the next contig with lower read depth in the series.

2.2.3.3 Scaffolding

The deduplicated assembly from Canu was polished with the PacBio Sequel II HiFi CCS reads using two iterations of `Racon v1.4.3` (Vaser et al. 2017). Hybrid scaffolds were generated using the *de novo* Bionano maps and the polished assembly (<https://bionanogenomics.com/support-page/data-analysis-documentation/>). To further resolve repetitive regions and improve assembly contiguity, the bionano-scaffolded assembly was integrated into the HERA pipeline (Du and Liang 2019). The Hi-C sequencing data was aligned with `bwa mem v0.7.17` (Li and Durbin 2009), PCR duplicates marked with `picard tools v1.83` (<http://broadinstitute.github.io/picard>) and the quality assessed with the `hic_qc.py` tool of Phase Genomics (https://github.com/phasegenomics/hic_qc). The assembly was then further scaffolded with the Hi-C alignments using `SALSA v2.2` (Ghurye et al. 2017), and subsequently polished with the PCR-free Illumina data using two iterations of `Pilon v1.23` (Walker et al. 2014). The final assembly at this point was the result of a meta-assembly with `quickmerge v0.3` (Chakraborty et al. 2016), which combined this version with an earlier draft version assembled directly from the PacBio CCS reads using Canu 1.8 (Koren et al. 2017), polished only with the Illumina PCR-free short-reads, but otherwise following an almost identical workflow. This was done in order to help mitigate the

possibility of misassembly arising from technical sources and thus improve overall contiguity. This resulting assembly was evaluated with BUSCO (Simão et al. 2015) and QAST v5.0.2 (Gurevich et al. 2013). Intermediate assembly statistics are given in comparison to i) immediately after Canu, and ii) the final version after re-scaffolding, in Suppl. Table A.2.

2.2.4 Re-scaffolding

Initial inspection of both gene and TE distributions and methylation patterns (methods described in Chapter 3) pointed to potential misassemblies in the assembled genome. Further investigation by way of synteny comparison to a closely-related species, *Eutrema salsugineum* (Yang et al. 2013), revealed that several of these likely occurred during scaffolding as orientation errors. Some of these errors could also be supported by comparison to the recently assembled Chinese accession (YUN_Tarv1.0) of *T. arvense* (Geng et al. 2021).

2.2.4.1 Development of genetic maps

Two genetic linkage maps were obtained using F2 populations in order to help correct misassemblies and help improve overall contiguity. The first linkage map was derived from a cross between MN106-Ref and a genetically distant Armenian accession Ames32867. The resulting F1 plants were allowed to self-fertilise, and seeds from a single plant were collected and propagated to the F2 generation. Approximately 500 mg fresh tissue was collected from 94 individuals in the F2 population. The tissue was desiccated using silica beads and pulverised using a tissue lyser. DNA was isolated with the Biosprint DNA plant kit (Qiagen, Valencia, CA, USA). The F2 population along with the two parental genotypes was genotyped via genotyping-by-sequencing at the University of Minnesota Genomics Centre (Minneapolis, MN, USA). Each sample was digested with the BtgI_BgLII restriction enzyme combination, barcoded, and sequenced on the Illumina NovaSeq S1 (Single-end 101 bp) yielding a mean of 1,237,890 reads per sample. The raw reads were de-multiplexed based on the barcode information and aligned to the most recent iteration of the pennycress genome using bwa. Sequence aligned files were processed with samtools v1.9 (Li et al. 2009) and picard tools to sort files and remove group identifiers. Variants were called using GATK HaplotypeCaller v3.3.0 (McKenna et al. 2010; Poplin et al. 2018). SNPs identified among these 94 lines were used for the development of genetic maps. The second linkage map was derived from a cross between MN106-Ref and a mutant line 2019-M2-111. To identify the variant alleles in 2019-M2-111, whole genome re-sequencing using paired-end reads was performed on the Illumina platform. SNPs were identified using a similar approach as described above. Sixty-seven SNP markers were designed using the biallelic information from re-sequencing data. DNA was

2 Building a Suitable Reference Genome

extracted from 48 samples from the mutant F2 population using the Sigma–Aldrich ready extract method, allele-specific and flanking primers synthesised from IDT (Iowa, USA) for each of the alleles were mixed, and genotyping was performed using the methods described in (Chopra, Folstad, and Marks 2020).

A total of 35,436 SNPs were identified among the population used for the first linkage map. SNP sites were selected for no-missing data, $QD > 1000$ and the segregation of the markers was 1:2:1, resulting in 743 high-quality SNPs retained for further analysis. A genetic map for the population was constructed using JoinMap 5 (Stam 1993). Only biallelic SNPs were used in the analysis and genetic maps were constructed with regression mapping based on default parameters of recombination frequency of < 0.4 with only the first two steps. The Kosambi mapping function was chosen for map distance estimation, and the Ripple function was deployed to confirm marker order within each of the seven linkage groups. A total of 319 markers were mapped to seven linkage groups. Similarly, 67 markers were genotyped on 48 individuals from the second population of linkage and 52 markers were mapped to six linkage groups. Both of these linkage maps were used in reordering and correcting the existing scaffolds as described below.

2.2.4.2 Re-scaffolding

In order to address misassemblies in the genome, several breakpoints were introduced manually at selected loci in the assembled genome where supported by at least two sources of data from: newly-derived genetic linkage maps (wild and EMS mutation based), Hi-C contact maps from the original library, whole genome alignments to accession YUN_Tarv1.0, and synteny maps to *E. salsugineum* (derived from reciprocal-best blast). These were cross-examined with minimap2 alignments of original PacBio CLR reads to the genome, an overview of corresponding gene and TE distributions produced by Liftoff v1.5.2 (Shumate and Salzberg 2020), and further comparative genomics with *E. salsugineum*. The resulting contigs were then re-scaffolded with ALLMAPS v1.1.5 (Tang et al. 2015), to produce the final, completed assembly by integrating both the synteny map and genetic map data and manually discounting contigs that were supported only by single markers. The final assembly statistics in comparison to previous intermediary stages are given in Suppl. Table A.2.

2.2.5 Comparative genomics

2.2.5.1 Genome sequences

Arabidopsis thaliana (Araport 11), *Schrenkiella parvula* (v2.2) and *Arabidopsis lyrata* (v2.1) genome sequences and gene annotation were downloaded from Phytozome (Goodstein et al. 2012). The *E. salsugineum* gene annotation was obtained from Phytozome and lifted over to the assembly obtained from NCBI (GCA_000325905.2).

2.2.5.2 Genome alignments and synteny analysis

Genome alignments between the initial and corrected versions of the *T. arvense* assembly to *E. salsugineum* were carried out using MUMmer v4.0.0 (Marçais et al. 2018), with a minimal length of 200 nt, followed by filtering for 1:1 matches and removing alignments smaller than 1000 bp. MCScan, a tool in the JCVI utility library (<https://github.com/tanghaibao/jcvi>) (Tang et al. 2008), was used to identify the interspecies gene orthologs and syntenic relationships between *T. arvense* and comparison species. The ortholog relationships were obtained using the proteinic translation of the CDS and using the parameter `--cscore=0.99`. To define the syntenic blocks and the corresponding genomic coordinates, the parameters `--minspan=15` and `--minsize=5` were also used. The genomic coordinates from the syntenic blocks were parsed to draw the syntenic relationships using Circos v0.69-8 (Krzywinski et al. 2009).

A set of ancestral chromosomal blocks (ABKs) defined by (Murat et al. 2015) provides a reference point by which to understand genomic rearrangement in the context of Brassicaceae evolution. To facilitate this comparison in pennycress, the ortholog relationships between each gene in *T. arvense* and *A. thaliana* obtained in the synteny analysis were cross-referenced with a published gene list where each ortholog gene of *A. thaliana* had an assigned ABK block (Murat et al. 2015).

2.3 Results

2.3.1 An improved reference genome sequence

The genome of *T. arvense* line MN106-Ref was assembled *de novo* from 476X (256 Gb) depth PacBio Sequel II CLR reads (~38 kb subread N50). The initial assembly attempts exceeded the genome size by ~53% with respect to the range of 459-539 Mbp total size estimated from flow cytometry and k-mer analysis (Suppl. Table A.1). Reducing the duplicated fraction, polishing, and scaffolding/re-scaffolding using a culmination of various approaches resulted in a final assembly

2 Building a Suitable Reference Genome

Table 1. Full descriptive statistics comparing the previously published T_arvense_v1 assembly to the present version T_arvense_v2.

Assembly category	T_arvense_v1	T_arvense_v2
# contigs	44,109	4,714
Largest contig	-	41.6 Mbp
contig N50	0.02 Mbp	13.3 Mbp
# scaffolds	6,768	964
# scaffolds ($\geq 50,000$ bp)	1,807	607
Largest scaffold	2.4 Mbp	70.0 Mbp
Total length	343 Mbp	526 Mbp
Total length ($\geq 50,000$ bp)	276 Mbp	514 Mbp
GC (%)	37.99	38.39
N50	0.14 Mbp	64.9 Mbp
NG50	0.05 Mbp	64.9 Mbp
N75	0.06 Mbp	61.0 Mbp
NG75	-	55.2 Mbp
L50	561	4
LG50	1,678	4
L75	1,469	6
LG75	-	7
# N's per 100 Kbp	5,165.00	0.51

of ~526 Mbp, corresponding to ~97.5% of the upper limit of the flow cytometry-based estimate and representing an improvement of ~20% relative to the original assembly size. Scaffolding/re-scaffolding of the genome assembly was achieved in the end using a combination of Bionano optical, Hi-C contact, genetic linkage and comparative synteny maps. The final genome contains 964 scaffolds, with ~83.6% of the total estimated size represented by seven large scaffolds, in agreement with the haploid chromosome number, demonstrating a vast improvement in overall contiguity in comparison to the previously-published draft assembly and bringing it to chromosome-level. The coding space is 98.7% complete on the basis of conserved core eukaryotic single-copy genes (BUSCO), with 92.1% being single-copy and 6.6% duplicated. Full descriptive statistics of the final version in comparison to T_arvense_v1 are given in Table 1; intermediary versions are summarised in Suppl. Table A.2.

In addition to aforementioned duplicate contigs arising from initial assembly (see methods), alignments of the raw CLR reads to the new genome revealed the presence of what appeared to be a small number of collapsed repeats in scaffolds 1, 3, 5, and 7. These are typically larger than 25

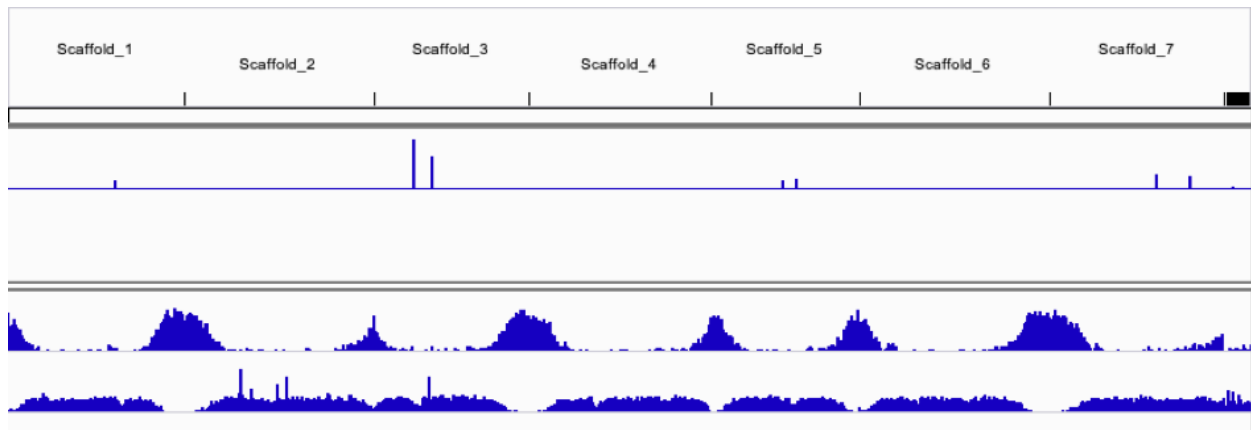


Figure 5. Integrative Genome Viewer (IGV) snapshot of PacBio read coverage (top track) over the largest seven scaffolds of the genome, including distributions of genes (middle track) and transposable elements (bottom track). Spikes in coverage in scaffolds 1, 3, 5, and 7 and indicative of collapsed repeats which are typically larger than the average read length.

Kbp and indicative of lingering misassemblies that remain unresolved in these loci (Figure 5). Further investigation revealed an overlap with tandem repeat clusters of 18S and 28S rRNA annotations at those loci on scaffolds 3 and 5, and a large supersatellite of 5S rRNA on scaffold 1. In addition, there were corresponding genes associated with organellar DNA at those loci on scaffolds 3 and 7, indicating either erroneous incorporation of plastome sequence during assembly or genuine nuclear integrations of plastid DNA (NUPT's) (Michalovova, Vyskot, and Kejnovsky 2013).

2.3.2 Comparative genomics

Exploiting information from the genome of *Eutrema salsugineum* (Yang et al. 2013), a closely related species (Franzke et al. 2011) with a much smaller genome (241 Mbp) but the same karyotype ($n=7$), aided during re-scaffolding (see methods; Suppl. Figure A.3) and confirmed the synteny of the seven largest scaffolds between the two species (Suppl. Figure A.4). There is a high level synteny between the two genomes, with the exception of some regions on scaffolds 2, 3, 6 and 7. This could be due to the low gene density observed in the *T. arvense* genome towards the centre of each chromosome (see Chapter 3) and/or the high presence of dispersed repeats in those regions.

Chromosome evolution in the Brassicaceae has been studied through chromosome painting techniques, and 24 chromosome blocks (A-X) have been defined from an ancestral karyotype of $n=8$ (Murat et al. 2015; Schranz, Lysak, and Mitchell-Olds 2006). These 24 blocks were identified in *T. arvense* based on gene homology and resulting synteny analysis between *T. arvense* and *A. thaliana* (Figure 6). While in general the distribution of the chromosomal blocks resembles that in

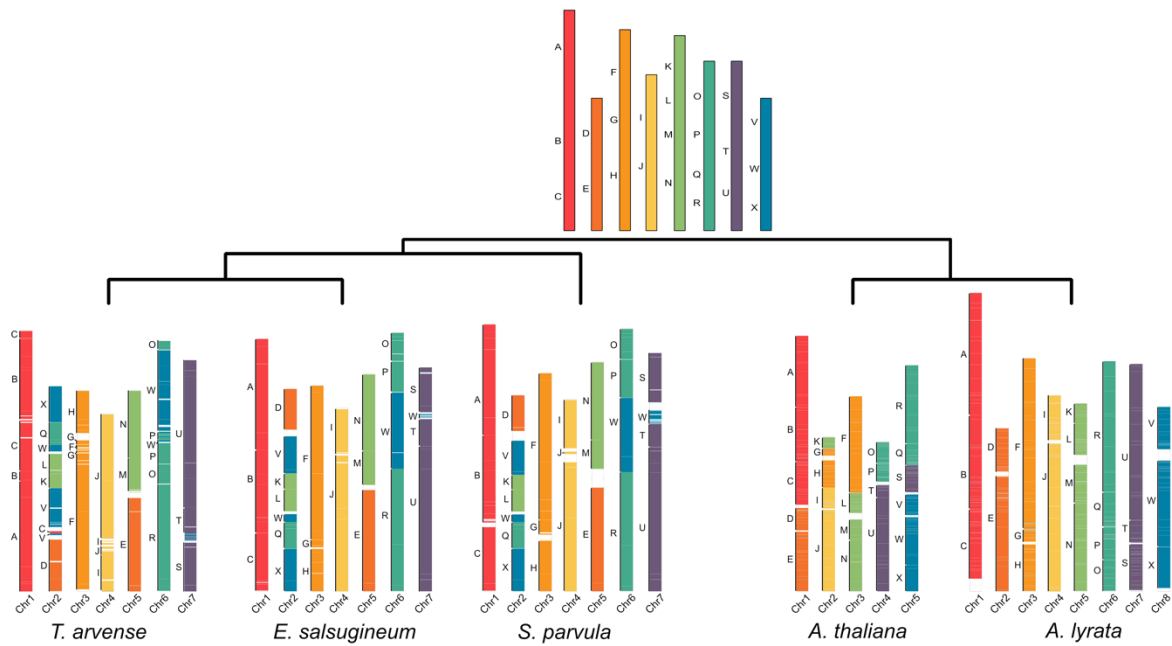


Figure 6. Distribution of ancestral genomic blocks (top panel) along the seven largest scaffolds of *T. arvense* MN106-Ref (T_arvense_v2), and a comparison of these genomic blocks with *Eutrema salsugineum*, *Schrenkiella parvula*, *Arabidopsis thaliana* and *Arabidopsis lyrata*.

the close relatives *E. salsugineum* and *S. parvula*, some blocks are rearranged in a small section at the end of the scaffold representing chromosome 1, and at the beginning of chromosome 6. The first case involves the transposition of a small part of block C in between A and B, while chromosome 6 has a possible inversion between the blocks O and W when compared to *E. salsugineum* and *S. parvula*. Overall, despite having an increase in genome size compared to *E. salsugineum* and *S. parvula*, *T. arvense* conserves all the ancestral Brassicaceae karyotype blocks. The synteny analysis also revealed intra-chromosomal rearrangements, but no obvious inter-chromosomal rearrangements.

2.4 Discussion

To facilitate robust biological inferences from genomic resources, it is important to have a reliable, high-quality genome assembly before further consideration can be given to downstream analyses. Over the last few years, significant efforts have been made towards the discovery of crucial traits and translational research in field pennycress, centering on MN106-Ref and the gene space information generated by (Dorn et al. 2013; 2015). Here, the reference genome has been re-assembled *de novo*, using a combination of PacBio CLR and CCS technology, alongside PCR-free Illumina, Bionano optical maps, Hi-C sequencing, and genetic maps. By leveraging the advantages of each technology, this hybrid approach resulted in a markedly improved level of quality and contiguity, as demonstrated in comparative genomics and in the estimation of k-mer complexity

and completeness. Alongside high-quality annotations, this improved resource helps make *T. arvense* line MN106-Ref more accessible both as a field-based model species for genetics and epigenetics studies, and to provide tools for its domestication as a new and extremely hardy winter annual cash cover crop.

Improved genomic resources help facilitate the general understanding of plant and evolutionary biology while also aiding plant breeding and crop improvement (Scheben, Yuan, and Edwards 2016). For example, pennycress and Arabidopsis share many of the key features that made Arabidopsis the most widely studied model plant system (Meinke et al. 1998). The use of Arabidopsis for translational research and for identifying potential gene targets in *T. arvense* is possible and has been extensively validated (Chopra et al. 2018, 2020; Chopra, Folstad, and Marks 2020; Jarvis et al. 2021; McGinn et al. 2019). Previous studies have suggested that over a thousand unique genes in *T. arvense* are represented by multiple genes in Arabidopsis and vice versa. Comparative genomics by way of synteny with *E. salicagineum* (Yang et al. 2013) revealed a high level of agreement, particularly between the protein-coding fraction of the genome, represented as conserved blocks in the largest seven scaffolds relative to the ancestral karyotype in Brassicaceae (Murat et al. 2015) (Figure 6). The detailed description of gene synteny between *T. arvense* and other Brassicaceae provides insights into its evolutionary relevance within lineage II. In addition, the difference in genome size between *T. arvense* and other species, despite the reduced level of gene duplication and the 1:1 gene relationship, can be explained by the large repetitive fraction present throughout both the centromeric and pericentromeric regions (see Chapter 3). In the absence of whole genome duplication events, this repetitive fraction indicates that the increased genome size may be a consequence of active TE expansion. This is therefore suggestive of a mechanism by which deleterious retrotransposon insertions must be mitigated in *T. arvense*. This could be explained by the high proportion of Gypsy retrotransposons in this species, usually located in heterochromatic regions, or by integration site selection (Sultana et al. 2017), or otherwise via silencing by small RNA activity and/or DNA methylation (Bucher, Reinders, and Mirouze 2012; Sigman and Slotkin 2016). Given the relatively high error rate of PacBio CLR reads (~10-15% before correction) with respect to circular consensus sequencing (CCS), the repetitive fraction would also help to explain the initial overestimation of the assembly size as a result of duplicated contigs. Several loci with highly overrepresented read coverage were also detected, indicative of repeat collapsing during the assembly process, often observed to be intersecting with 5S, 18S, and 28S rRNA annotations. Such regions are difficult even for current long read technologies due to the large size of the tandem repeat units.

2 Building a Suitable Reference Genome

To fully explore the extent of genomic variability among the population, the assembly of additional accessions can also help to further enrich the resources available for the study of pennycress. In parallel to this study, a Chinese accession of *T. arvense* (YUN_Tarv_1.0) was assembled using Oxford Nanopore, Illumina HiSeq and Hi-C sequencing (Geng et al. 2021). This timely availability of an additional frame of reference also opens the door to a pan-genomic approach in evolutionary research, allowing for the better characterisation of structural variants, for example. Furthermore, the use of different sequencing technologies and assembly software provides an additional avenue to correct misassemblies and base calling errors in either case. The overall longer contigs assembled with PacBio CLR, for example, and the consideration of various genetic map data in addition to Hi-C, provides a greater resolution of scaffolds particularly throughout the centromere and pericentromeric regions (Suppl. Figure A.5). The reduced error-rate of PacBio CCS (used for polishing) is also reflected in the overall k-mer content, which is measured with a two-order magnitude higher consensus quality over scaffolds representing chromosomes and ~99% overall completeness for T_arvense_v2 (Suppl. Tables A.4-6), indicative of high-quality, error-free sequences more appropriate for variant calling, for instance. Nevertheless, combination of these resources with those of (Geng et al. 2021) and other, similar initiatives will allow for the investigation of such differences that might exist between accessions originating from different geographic locations around the world, helping to provide further insight into both structural variations and evolutionary dynamics.

3 Feature Annotation for Epigenomics

3.1 Introduction

While a high-quality reference genome is often necessary to make an accurate assessment in regards to genetic and epigenetic variability, it would not be complete without a complementary annotation by which to understand the relative biological consequences. Within the context of epigenomics, these consequences can be reflected in terms of gene expression (Jaenisch and Bird 2003; Xiaoyu Zhang et al. 2006; Lang et al. 2017), the activation and silencing of transposable elements (Mirouze et al. 2009; Tsukahara et al. 2009), and in the activity of small RNA (Aufsatz et al. 2002; Cao et al. 2003; Matzke and Mosher 2014; Lei et al. 2015). A robust annotation which includes genes, TEs and the genomic loci associated with populations of small RNA are therefore of key interest during downstream analysis. Characterisation of tissue-specificity in terms of both a gene expression atlas and the DNA methylome also provides a frame of reference which is useful for both epigenetic research and to provide targets for genetic (or epigenetic) manipulation of field pennycress, in its placement as a newly-domesticated crop species. Herein, robust annotations are described for both protein-coding and non-coding genes, including putative transfer RNA (tRNA), ribosomal RNA (rRNA) and small nucleolar RNA (snoRNA) predictions, alongside small RNA (sRNA) producing loci, transposable element (TE) families, and predicted pseudogenes. The gene expression atlas is built from transcriptome data based on a panel of eleven different tissues and life stages, which in combination with whole-genome DNA methylation profiles of both roots and shoots provides a basis for exploring gene regulatory and/or epigenetic mechanisms within pennycress. The genome and resequencing information presented in this study will increase the value of pennycress as both a model and a tool for translational research, and accelerate pennycress breeding through the discovery of genes affecting important agronomic traits.

3.2 Materials and methods

3.2.1 Tissue preparation for RNA sequencing

MN106-Ref seeds were surface-sterilised with chlorine gas for 1 h and stratified for 3 d at 4°C. For seedling-stage RNA extractions, seeds were plated on ½ MS medium supplement with 1% plant

3 Feature Annotation for Epigenomics

Agar and stratified for 3 d at 4°C. For all other tissue collections, plants were sown on soil and grown in a climate-controlled growth chamber in long-day conditions (16/8 h light/dark at 21°/16°C, light intensity 140 $\mu\text{E} / \text{m}^2\text{s}$, with 60% relative humidity; plants were watered twice per week. Two weeks after germination, plants growing on soil were vernalised at 4°C in the dark for 4 weeks, then moved back to the growth chamber. Samples were collected from 11 different tissues, each with three biological replicates (two in case of mature seeds); for each replicate, tissue was pooled from two individuals. Tissues included: one-week old shoots (from plate culture), one-week old roots (from plate culture), rosette leaves, cauline leaves, inflorescences, open flowers, young green siliques ($\sim 0.5 \times 0.5$ cm), older green siliques ($\sim 1 \times 1$ cm), seed pods, green seeds, and mature seeds.

3.2.2 RNA extraction and sequencing

Total mRNA was extracted using the RNeasy Plant Kit (Qiagen, Valencia, CA, USA) and treated with DNase I using the DNA-free kit DNase Treatment and Removal Reagents (Ambion by life technologies, Carlsbad, CA, USA), following recommended manufacturer protocols. cDNA libraries were constructed using the NEBNext Ultra II Directional RNA Library Prep kit (New England BioLabs, Ipswich, MA, USA inc.) for Illumina following the manufacturer's protocol. Libraries were sequenced on a HiSeq 2500 instrument (Illumina, San Diego, CA, USA) as 125 bp paired-end reads.

3.2.3 Transcriptome assembly

Following quality control and adapter clipping with cutadapt v2.6 (M. Martin 2011), biological replicates for each of eleven tissue types from Illumina mRNA-seq libraries were aligned independently using STAR v2.5.3a (Dobin et al. 2013), then merged according to tissue type, prior to assembly via a reference-based approach. Each assembly was performed using Ryuto v1.3m (Gatter and Stadler 2019), and consensus reconstruction was then performed using TACO v0.7.3 (Niknafs et al. 2017) to merge tissue-specific transcriptome assemblies. PacBio Iso-seq libraries from MN106-Ref were refined, clustered and polished following the Iso-seq3 pipeline (<https://github.com/PacificBiosciences/IsoSeq>), prior to alignment with STARlong and isoform collapsing using the cDNA_Cupcake (https://github.com/Magdoll/cDNA_Cupcake) suite. The Iso-seq data was later leveraged together with the Illumina mRNA-seq data to prioritise convergent isoforms using custom in-house scripting.

3.2.4 Genome annotation

The newly-assembled genome for MN106-Ref was annotated using the MAKER-P v2.31.10 (Campbell, Holt, et al. 2014; Campbell, Law, et al. 2014) pipeline on the servers provided by the EpiDiverse project, at ecSeq Bioinformatics GmbH (Leipzig, Germany). Plant proteins were obtained from the Viridiplantae fraction of UniProtKB/Swiss-Prot and combined with RefSeq sequences derived from selected Brassicaceae: *Arabidopsis thaliana*, *Brassica napus*, *Brassica rapa*, *Camelina sativa* and *Raphanus sativus*. TEs were obtained from RepetDB (Amselem et al. 2019) for selected plant species: *Arabidopsis lyrata*, *Arabidopsis thaliana*, *Arabis alpina*, *Brassica rapa*, *Capsella rubella*, and *Schrenkiella parvula* (*Eutrema parvulum*). Repeat library construction was carried out using RepeatModeler v1.0.11 (Smit and Hubley 2008) following basic recommendations from MAKER-P (Campbell, Holt, et al. 2014). Putative gene fragments were filtered out following BLAST search to the combined Swiss-Prot + RefSeq protein plant database after exclusion of hits from RepetDB. The *de novo* library was combined with a manually curated library of plant sequences derived from rebase (Bao, Kojima, and Kohany 2015). Genome masking is performed with RepeatMasker v4.0.9 (Smit 2004) as part of the MAKER-P pipeline. Protein-coding genes, non-coding RNAs and pseudogenes were annotated with the MAKER-P pipeline following two iterative rounds under default settings, using (i) transcript isoforms from Illumina mRNA-seq and PacBio Iso-seq data, (ii) protein homology evidence from the custom Swiss-Prot + RefSeq plant protein database, and (iii) the repeat library and TE sequences for masking. The initial results were used to train gene models for ab initio predictors SNAP v2006-07-28 (Korf 2004) and Augustus v3.3.3 (Stanke et al. 2006), which were fed back into the pipeline for the subsequent rounds. The final set of annotations were filtered based on Annotation Edit Distance (AED) < 1 except in cases with corresponding PFAM domains, as derived from InterProScan v5.45-80.0 (Jones et al. 2014). The tRNA annotation was performed with tRNAscan-SE v1.3.1 (Lowe and Eddy 1997) and the rRNA annotation with RNAmmer v1.2 (Lagesen et al. 2007), respectively. The snoRNA homologs were derived using Infernal v1.1.4 (Nawrocki and Eddy 2013) from plant snoRNA families described in (Patra Bhattacharya et al. 2016). A small phylogeny based on gene orthologs and duplication events in comparison to selected Brassicaceae (*A. lyrata*, *A. thaliana*, *B. rapa*, *S. parvula*, and *E. salsugineum*) was performed with OrthoFinder v2.5.2, and the resulting species tree is rooted using STRIDE (David M. Emms and Kelly 2017) and inferred from all genes using STAG (David M. Emms and Kelly 2018).

3.2.5 Transposable element annotations

Two *de novo* annotation tools, EDTA v1.7.0 (Ou et al. 2019) and RepeatModeler v2.0 (Flynn et al. 2020), were used to annotate TEs independently. For EDTA the following parameters were used in addition to defaults: `--species others`, `--step all`, `--sensitive 1`, `--anno 1`, and `--evaluate 1`. For RepeatModeler2 the additional parameters were `-engine ncbi` and `-LTRStruct`. The outputs of both tools were evaluated by manual curation. First, tblastn was used to align each TE consensus with the transposase database obtained from rebase, and the retrotransposon domains (GAG, Pol, Env, etc.) were viewed in turn using dotter (Sonnhammer and Durbin 1995). Sequences with multiple paralogs were mapped back to the genome and manually extended to determine the full length boundary of each TE. A total of 107 full length, representative Copia and Gypsy families were successfully evaluated. The TE consensus from RepeatModeler2 was selected as the most accurate model, based on full length paralogs. RepeatMasker was then used to construct the GFF3-like file from the FASTA file from RepeatModeler2, with the optional settings: `-e ncbi -q -no_is -norna -nolow -div 40 -cutoff 225`. The perl script `rmOutToGFF3.pl` was used to generate the final GFF3 file.

3.2.6 Small RNA annotations

3.2.6.1 sRNA plant material.

Seeds were sterilised by overnight incubation at -80°C , followed by 4 hours of bleach treatment at room temperature (seeds in open 2 mL tube in a desiccator containing a beaker with 40 mL chlorine-based bleach (<5%; DanKlorix, Colgate-Palmolive, New York, NY, USA) and 1 mL HCl (32%; Carl Roth, Karlsruhe, Germany)). For rosette, inflorescence and pollen samples, seeds were stratified in the dark at 4°C for six days prior to planting on soil, then cultivated under growth chamber conditions of 16-23 $^{\circ}\text{C}$, 65% relative humidity, and a light-dark photoperiod of 16h:8h under 110-140 $\mu\text{mol m}^{-2} \text{s}^{-1}$ light. Rosette leaves were harvested after two weeks of growth. For inflorescence and pollen, six week old plants were vernalised for four weeks at 4°C in a light-dark photoperiod of 12h:12h under 110-140 $\mu\text{mol m}^{-2} \text{s}^{-1}$ light. Two weeks after bolting, inflorescence and pollen was collected. Pollen grains were collected by vortexing open flowers in 18% sucrose for 5 min followed by centrifugation at 3,000g for 3 min in a swinging bucket rotor. For root samples, seeds were stratified for six days at 4°C in the dark on $\frac{1}{2}$ MS media. Plants were grown in 3-4 mL $\frac{1}{2}$ MS medium plates in long-day (16 hours) at 16°C . Root samples were collected 12-14 days after stratification.

3.2.6.2 sRNA extraction and library preparation

Total RNA was extracted by freezing collected samples with liquid nitrogen and grinding with a mortar and pestle with Trizol reagent (Life Technologies, Carlsbad, CA, USA). Then, total RNA (1 µg) was treated with DNase I (Thermo Fisher Scientific, Waltham, MA, USA) and used for library preparation. Small RNA libraries were prepared as indicated by the TruSeq small RNA library prep kit (Illumina, San Diego, CA, USA), using 1 µg of total RNA as input, as described by the TruSeq RNA sample prep V2 guide (Illumina, San Diego, CA, USA). Size selection was performed using the BluePippin System (SAGE Science, Massachusetts, USA). Single-end sequencing was performed on a HiSeq3000 instrument (Illumina, San Diego, CA, USA).

3.2.6.3 sRNA loci annotation

Raw FASTQ files were processed to remove the 3'-adapter and quality controlled with `trim_galore v0.6.6`

(https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) using `trim_galore -q 30 --small_rna`. Read quality was checked with `FastQC v0.11.9` (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>). The reference annotation of sRNA loci was created following the steps indicated by (Lunardon et al. 2020). In short, each library was aligned to the reference genome independently using `ShortStack v3.8.5` (Axtell 2013b), with default parameters, to identify clusters of sRNAs *de novo* with a minimum expression threshold of 2 reads per million (RPM). sRNA clusters from all libraries of the same tissue were intersected using `BEDTools v2.26.0 multiIntersectBed` (Quinlan and Hall 2010) with default parameters, and only those loci present in at least three libraries were retained. For each tissue, sRNA clusters 25 nt apart were padded together with the `bedtools merge -d` option. sRNA loci with an expression of <0.5 RPM in all libraries of each tissue were also removed. Finally, sRNA loci for all different tissues were merged in a single file retaining tissue of origin information with `bedtools merge -o distinct` options. miRNAs predicted by the `ShortStack` tool were manually curated following the criteria of Axtell (2013a): Maximum hairpin length of 300 nt; $\geq 75\%$ of reads mapping to the hairpin must belong to the miRNA/miRNA* duplex; For the miRNA/miRNA* duplex: No internal loops allowed, two-nucleotide 3' overhangs, maximum five mismatched bases, only three of which are nucleotides in asymmetric bulges; Mature miRNA sequence should be between 20 nt and 24 nt.

3.2.7 Expression atlas

Gene expression was measured from the same tissue-specific STAR alignments taken prior to merging biological replicates for transcript assembly, excluding coverage outliers “mature seed” and “green old silique”. A total of 27 samples from 9 tissues were therefore considered for gene expression analysis. Raw counts were generated using subread featureCounts v2.0.1 (Liao, Smyth, and Shi 2014) and subsequently normalised using the trimmed mean of M-values (TMM) (Robinson and Oshlack 2010) derived from edgeR v3.34 (Robinson, McCarthy, and Smyth 2010). Averaged expression counts by group were taken for tissue specificity evaluation using the Tau (τ) algorithm (Yanai et al. 2005), as implemented in the R package `tispec` v0.99.0 (<https://rdrr.io/github/roonysgalbi/tispec/>), which provides a measure of τ in the range of 0-1 where 0 is non/low specificity and 1 indicates high/absolute specificity.

3.2.8 DNA methylation

Genomic DNA was extracted from roots and shoots of 2-week-old seedlings grown on $\frac{1}{2}$ MS medium with 0.8% Agar and 0.1% DMSO. Seedlings were grown vertically in 16h/8h light/dark; at the time of sampling, roots were separated from shoot tissue with a razor blade and the plant tissue was flash-frozen in liquid nitrogen. Genomic DNA was extracted from ground tissue using the DNeasy Plant Mini kit (QIAGEN, Hilden, Germany). Libraries for WGBS were prepared using the NEBNext UltraII DNA Library Prep kit (New England Biolabs). Adapter-ligated DNA was treated with sodium bisulfite using the EpiTect Plus Bisulfite kit (QIAGEN, Hilden, Germany) and amplified using the Kapa HiFi Uracil + ReadyMix (Roche, Basel, Switzerland) in 10 PCR cycles. WGBS libraries were sequenced on an Illumina HiSeq2500 instrument with 125 bp paired-end reads.

The WGBS libraries were processed using the `nf-core/methylseq` v1.5 pipeline (10.5281/zenodo.2555454), combining `bwa-meth` v0.2.2 (Pedersen et al. 2014) as an aligner and `MethylDackel` v0.5.0 (<https://github.com/dpryan79/MethylDackel>) for the methylation calling. The default parameters were used for the entire workflow, with the exception of the methylation calling where the following arguments were used: `-D 1000 --maxVariantFrac 0.4 --minOppositeDepth 5 --CHG --CHH --nOT 3,3,3,3 --nOB 3,3,3,3 -d 3`. Only cytosines with a minimum coverage of 3X were kept for the subsequent analysis. Further comparisons between the methylated cytosines and the genome annotation were performed using `BEDTools` v2.27.1 (Quinlan and Hall 2010).

3.3 Results

3.3.1 Transcriptome assembly

Total cDNA was sequenced using strand-specific RNA-seq from eleven tissues, including rosette leaves, cauline leaves, inflorescences, open flowers, young green siliques, old green siliques, green seeds, mature seeds, seed pods, roots of 1-week-old seedlings, and shoots of 1-week-old seedlings (Suppl. Table B.1). Reads from each tissue sample were aligned to the newly-assembled MN106-Ref genome resulting in unique mapping rates in the range of ~76-91%, with the exception of old green silique (19%), green seed (59%), and mature seed (12%). The majority of unmapped reads in each case were due to insufficient high-quality read lengths. Independent, tissue-specific transcriptome assemblies were constructed and combined into a multi-sample *de novo* assembly, yielding 30,650 consensus transcripts. These were further refined by prioritising isoforms supported by Iso-seq data, resulting in 22,124 high-quality consensus transcripts taken forward to inform gene models.

3.3.2 Protein-coding genes

In addition to the expression data, the gene models were informed by protein homology using a combined database of Viridiplantae from UniProtKB/Swiss-Prot (Boutet et al. 2007) and selected Brassicaceae from RefSeq (Pruitt et al. 2012). Following initial training and annotation by *ab initio* gene predictors, protein-coding loci were further annotated with InterPro to provide PFAM domains, which were combined with a BLAST search to the UniProtKB/Swiss-Prot Viridiplantae database to infer gene ontology (GO) terms. In accordance with MAKER-P recommendations (Campbell, Holt, et al. 2014), the final set of 27,128 protein-coding loci were obtained by filtering out those with an annotation edit distance (AED) score of 1 unless they also contained a PFAM domain. Approximately 95% of loci had an AED score < 0.5 (Suppl. Figure B.1), demonstrating a high level of support with the available evidence, and 21,171 (~78%) were also annotated with a PFAM domain. Analysis of gene orthologs and paralogs among related Brassicaceae confirmed the close relationship with *E. salsugineum*, with the protein-coding fraction occupying a genome space comparable to related species (Figure 7a). A total of 4,433 gene duplication events were recorded with OrthoFinder, comparable to *E. salsugineum* (5,108), but fewer than in *B. rapa* (11,513), for example.

Full descriptive statistics of the new annotation are given in Table 2, in comparison to the original T_arvense_v1 annotation (Dorn et al. 2015) lifted over to the new genome with Liftoff v1.5.2

3 Feature Annotation for Epigenomics

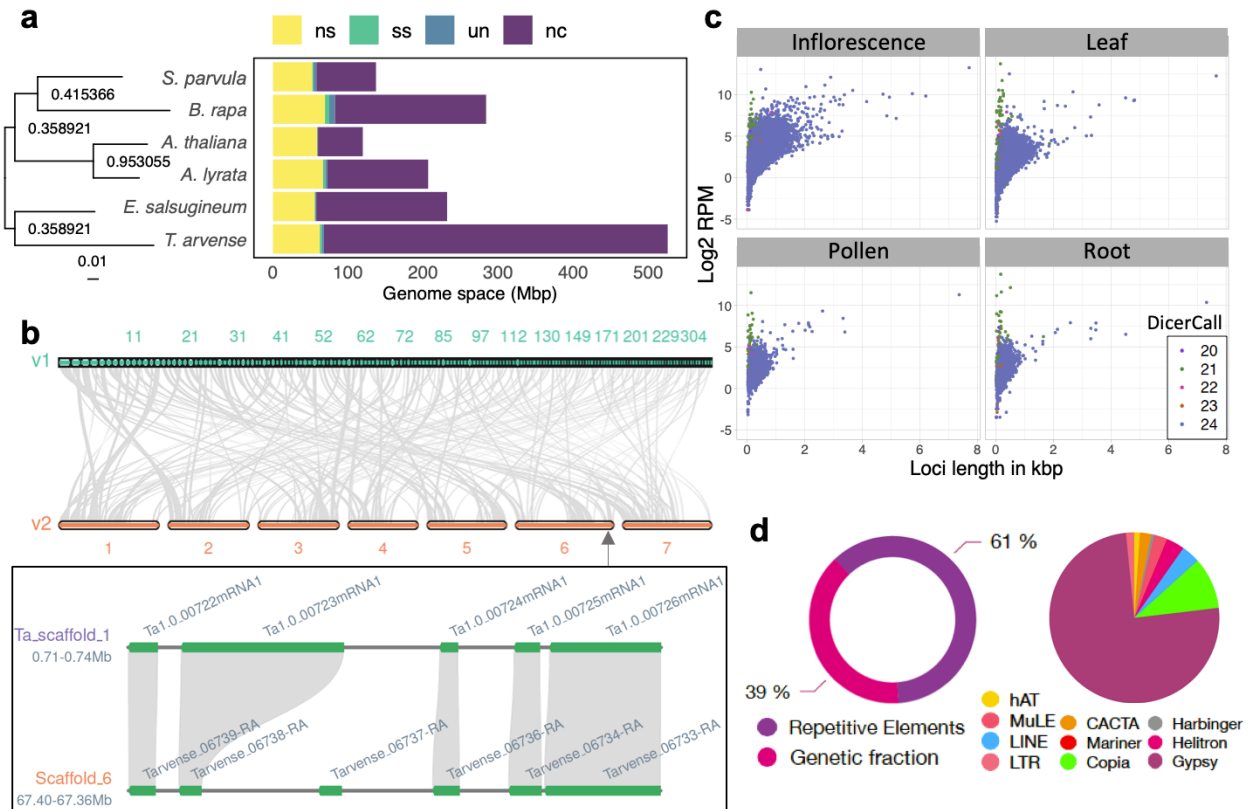


Figure 7. Feature annotations within *T. arvense* var. MN106-Ref. **a)** Rooted species tree inferred from all genes, denoting node support and branch length in substitutions per site, and horizontal stacked bar chart comparing the genetic fraction in pennycress with other Brassicaceae sp. (ns = non-specific orthologs, ss = species-specific orthologs, un = unclassified genes, nc = non-coding/intergenic fraction). **b)** Comparison of gene macrosynteny between v1 and v2 of the genome, and a microsynteny example of genes *MYB29* and *MYB76*, which are resolved in the v2 annotation. **c)** Small RNA biogenesis loci length and expression values in each of four tissues. **d)** Overall repetitive content in the genome as discovered by RepeatMasker2, and relative abundance of TEs within the fraction of repetitive elements.

(Shumate and Salzberg 2020), where applicable. Gene feature distributions are comparable between T_arvense_v1 and the present assembly of MN106-Ref (hereafter referred to as T_arvense_v2) (Suppl. Figure B.2). Unique genes that were successfully lifted over from the previous version were included as a separate fraction in the final annotation (source: T_arvense_v1), resulting in 32,010 annotated genes in total. Up to ~95.2% completeness can be obtained by combining the full set of both the current and previous annotations according to a BUSCO evaluation of 2121 conserved, single-copy orthologs. The improved contiguity of the genome space allowed for the resolution of genes such as the tandem duplicated *MYB29* and *MYB76*, which were concatenated in the previous version (Figure 7b).

Table 2. Summary of feature annotations in comparison to the original assembly version T_arvense_v1.

Type	T_arvense_v1	T_arvense_v2	diff.
(A) Protein-coding genes			
Total number of loci	27,390	27,128	-262
Total number of unique loci	4,780	5,034	+254
Total number of transcript isoforms	-	30,650	+30,650
Number of matching loci with changes in CDS	-	-	+14,102
Number of matching loci with changes in UTR(s)	-	-	+22,559
Loci containing one or more PFAM domain	-	21,171	+21,171
Loci annotated with one or more GO term	-	13,074	+13,074
(B) Non-coding genes			
tRNA	-	1,148	+1,148
rRNA clusters (<25 Kbp)	-	63	+63
snoRNA	-	243	+243
Small interfering RNA (siRNA)	-	19,373	+19,373
Micro RNA (miRNA)	-	71	+72
(C) Other gene types			
Pseudogenes (set II Ψs)	-	44,490	+44,490
Transposable element genes	-	423,251	+423,251

3.3.3 Non-coding loci

In addition to protein-coding genes, the annotation includes non-coding RNA (ncRNA) genes, pseudogenes, and TEs. Descriptive annotation statistics have been summarised in Table 2. While many of these annotation features in *T. arvense* were similar to those found in other plant species, several unique patterns were observed, which are described in detail below. ncRNA annotations were inferred from either sequence motifs (tRNA, rRNA, snoRNA) or from sequencing data (siRNA, miRNA), where appropriate. Clusters of both 5S rRNA and tandem repeat units of 18S and 28S rRNA were predicted with RNAmmer (Lagesen et al. 2007), which were often observed in relative proximity to loci identified with Tandem Repeats Finder v4.09.1 (Benson 1999) and putatively associated with centromeric repeat motifs (not shown). Of the largest seven scaffolds, only scaffolds 4 and 7 carried no such annotations. Notably, several large clusters of 5S rRNA genes were interspersed throughout the pericentromeric region of scaffold 1, whereas the remaining four scaffolds contained 18S and 28S rRNA gene annotations. Finally, 243 homologs were identified from 114 snoRNA families.

Table 3. Detailed per-class statistics of the transposable element fraction of the *T. arvense* genome.

Family	Key Name	Count	bp masked	% masked
hAT	DTA	7,449	3,312,483	0.63
CACTA	DTC	12,085	6,997,150	1.33
Harbinger	DTH	6,187	1,832,186	0.35
MuLE	DTM	18,022	8,017,253	1.53
Mariner	DTT	706	101,162	0.02
Helitron	DHH	24,151	11,129,635	2.12
LINE	RIC,RII,RIL,RIX	26,284	11,390,482	2.18
Copia	RLC	37,544	31,386,966	5.97
Gypsy	RLG	282,353	241,563,874	45.96
LTR	RLA	9,506	5,085,962	0.97

3.3.4 Transposable elements

In total, 423,251 TEs were identified, belonging to 10 superfamilies and covering ~61% of the genome (Figure 7d). Retrotransposons (75% of all TEs are Gypsy elements; 10% Copia; 4% LINE) by far outnumbered DNA transposons (3% Helitrons; 1% hAT; 2% CACTA; 1% Pif-Harbinger; 2% MuLE). A detailed breakdown of repeats can be found in Table 3. As the most abundant retrotransposon superfamily, Gypsy elements accounted for 46% of the total genome space, which is consistent with a high abundance observed in the pericentromeric heterochromatin of *E. sanguineum*, where centromere expansion is thought to have been caused by Gypsy proliferation (Zhang et al. 2020). In addition, we identified 359 protein-coding genes located fully within TE-bodies that could represent Pack-TYPE elements and contribute to gene shuffling (Catoni et al. 2019). Among these elements 153 were intersecting with mutator-like elements suggesting they correspond to Pack-MULE loci. TEs were located primarily in low gene density regions, while the fraction of TE-contained genes were randomly distributed.

3.3.5 Small RNA

In total, 19,386 siRNA loci were identified. More than 98% of these loci corresponded to heterochromatic 23-24 nt siRNA loci, with only 196 producing 20-22 nt siRNAs. The sRNA loci were expressed unevenly across tissues, as inferred from prediction with data from different tissues. Only 2,938 loci were shared across all four tissues studied (rosette leaves, roots, inflorescences, and pollen). Inflorescences were the major contributor with 6,728 private loci. Despite these

differences between tissues, we observed similar overall patterns in terms of locus length, expression (Figure 7c), and complexity (Suppl. Figure B.3).

Altogether, sRNA loci accounted for ~8 Mbp or ~1.5% of the assembled MN106-Ref genome. Of the seven largest scaffolds, where the majority of genes are located, the total coverage of siRNA loci ranged between 1.5 - 2% and the loci appeared to be preferentially concentrated at the boundary between TEs and the protein-coding gene fraction of the genome. In further exploration, the seven largest scaffolds were partitioned into gene-enriched and gene-depleted regions, based on a median of 14 genes per Mbp and a mean of 54.2 genes per Mbp. Gene-enriched loci were defined to be those above the mean, and gene-depleted loci as those below. At the (pseudo)chromosomal level, sRNA loci correlated with gene-enriched regions and were scarce in regions with high TE content. This trend is in contrast to that observed in *A. thaliana* (Hardcastle, Müller, and Baulcombe 2018) but resembles what has been observed, for example, in maize (He et al. 2013) and tomato (Tomato Genome Consortium 2012).

Phased secondary siRNAs (phasiRNAs) are a class of secondary sRNAs that, due to the way they are processed, produce a distinct periodical pattern of accumulation (Axtell 2013a). In the *T. arvense* genome, we observed 139 loci with such phased patterns. In contrast to the general notion that phasiRNAs are typically 21 nt long (Lunardon et al. 2020), we found 24 nt siRNAs to be dominant in 133 of these loci.

3.3.5.1 Micro RNAs

MicroRNA (miRNA)-encoding genes were predicted using a combination of ShortStack and manual curation (see Methods section 3.2.6). A total of 72 miRNA-producing loci were identified, including 53 that were already known from other species, and 19 which appeared to be species-specific. Most of the identified families were produced from only one or two loci; miR156 and miR166 were produced by the most loci, with eight and five family members, respectively. A total of 21 out of 25 families in *T. arvense* are found in other Rosids, and three (miR161, miR157, and miR165) only in other Brassicaceae. One family, miR817, is also present in rice. There is a strong preference for 5'-U at the start of both unique and conserved miRNAs (Suppl. Figure B.4), in line with previous reports (Voinnet 2009). The expression level of both conserved and novel miRNA families was compared between tissues, showing that the ten most highly-expressed across all tissues are conserved families whereas novel miRNA demonstrates a marginal tendency to be more lowly-expressed or with potential for differential expression (Suppl. Figure B.5).

3.3.5.2 sRNA genomic origin loci

When overlaying sRNA loci with the complete annotation of genes and TEs, most sRNAs localised to the intergenic space. A substantial fraction, especially 20-22 nt sRNAs, were however produced from intronic sequences (Suppl. Figure B.6a). Helitrons make up only 1.5% of the genome space, yet more than 5% of sRNA biogenesis loci overlap with this type of TE. Most sRNA loci (93.0%) fell within 1.5 kbp of annotated genes or TEs (Suppl. Figure B.6b,c). As expected, 23-24 nt sRNAs were more frequently associated with TEs, whereas 20-22 nt sRNAs more often produced by coding genes (Axtell 2013a).

3.3.6 Pseudogenes

In accordance with the MAKER-P protocol, pseudogenes (Ψ) were predicted in intergenic DNA with the ShiuLab pseudogene pipeline (Zou et al. 2009). A total of 44,490 set II pseudogenes were annotated, exceeding those in *A. thaliana* (~3,700) or rice (~7,900) by one order of magnitude. A total of 35,818 pseudogenes were observed overlapping with TEs, whereas 8,672 pseudogenes were either concentrated in intergenic space or more towards the protein-coding gene complement of the genome, and thus perhaps less likely to have arisen from retrotransposition. Approximately 59.2% of these contained neither a nonsense nor a frameshift mutation, indicating either (i) that the regulatory sequences of the pseudogenes were silenced first, (ii) a pseudo-exon which may be linked to another non-functional exon, or (iii) a possible undiscovered gene.

3.3.7 Gene expression atlas

Tissue-specific expression patterns could be elucidated from cDNA sequences arising from 11 different tissues or developmental stages. The complete expression atlas is provided in Nunn et al. (2022). The relative extent of tissue-specific gene expression was evaluated using the Tau (τ) algorithm (Yanai et al. 2005), from the normalised trimmed mean of M-value (TMM) counts in all tissues (Robinson and Oshlack 2010). To preclude potential biases caused by substantial differences in library size, low-coverage samples from mature seeds and old green siliques were excluded. In total, 4,045 genes had high or even complete tissue specificity ($\tau = 0.8 - 1.0$), whereas 5,938 genes had intermediate (0.2 - 0.8), and 6,107 had no or low specificity (0 - 0.2); the remaining genes were ignored due to missing data. The relative breakdown of each specificity fraction by tissue type is shown in Figure 8a, with “roots”, “green seeds”, and “inflorescences” representing the tissues with the greatest proportion of high or complete specificity genes. The relative

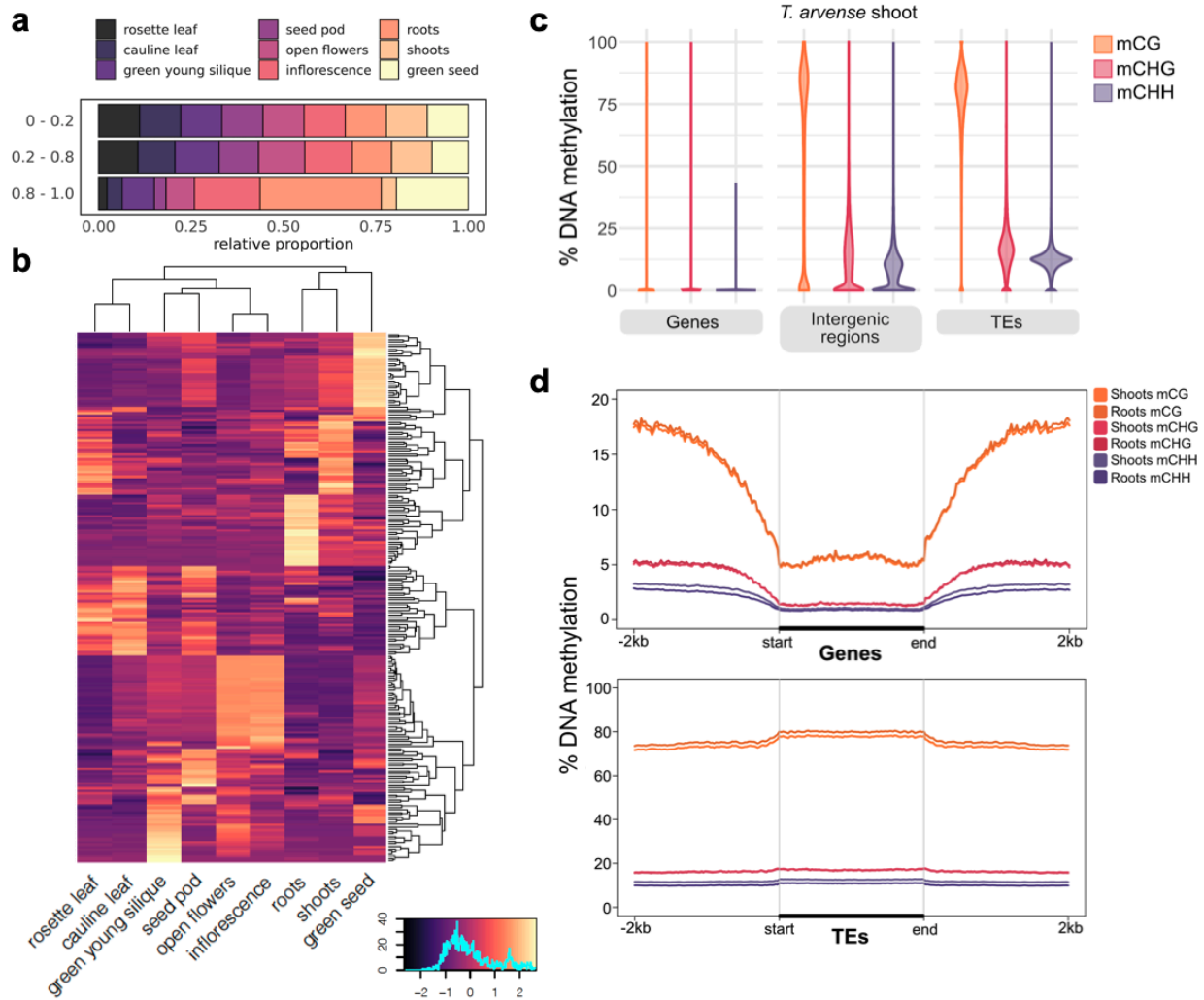


Figure 8. Regulatory dynamics in pennycress. **a)** Relative fraction of genes in each tissue for low (0 – 0.2), intermediate (0.2 – 0.8) and high/absolute specificity (0.8 – 1.0) subsets. **b)** $\log_2(\text{TMM})$ expression values of the top 30 most highly expressed genes in each tissue, relative to the mean across all tissues, from the subset of genes with a high/absolute tau specificity score. **c)** Distribution of average DNA methylation for different genomic features, by cytosine sequence context. **d)** DNA methylation along genes (top) and TEs (bottom), including a 2 kb flanking sequence upstream and downstream. DNA methylation was averaged in non-overlapping 25 bp windows.

$\log_2(\text{TMM})$ expression values of the top 30 most highly expressed genes in each tissue, given a high or complete specificity score, are plotted in Figure 8b with respect to the overall mean expression per gene across all included tissues. These include, for example, genes with homology to EXTENSIN 2 (EXT2; *A. thaliana*) in “roots”, CRUCIFERIN (BnC1; *B. napus*) in “green seeds”, and PECTINESTERASE INHIBITOR 1 (PMEI1; *A. thaliana*) in “inflorescences”, and “open flowers”.

3.3.8 DNA Methylation

Understanding the distribution of cytosine methylation in the MN106-Ref line can provide a basis for comparison in other studies which seek to use the reference genome as a resource, and can also help characterise its role in the genomic landscape. Previous studies in model species have shown it to be associated with heterochromatin and transcriptional inactivation of TEs and promoters, but also with higher and more stable expression when present in gene bodies (Zhang, Lang, and Zhu 2018). In plants, DNA methylation occurs in three cytosine contexts, CG, CHG, and CHH (where H is any base but G), with the combined presence of CG, CHG and CHH methylation usually indicative of heterochromatin formation and TE silencing, while gene body methylation consists only of CG methylation (Bewick and Schmitz 2017).

In light of the high TE density in *T. arvense*, genome-wide DNA methylation by whole-genome bisulfite sequencing (WGBS) was subsequently analysed in shoots and roots of 2-week-old seedlings. Genome wide, 70% of cytosines were methylated in the CG context, 47% in the CHG context, and 33% in the CHH context. In line with findings in other Brassicaceae, methylation at CG sites was consistently higher than at CHG and CHH (Figure 9a; Suppl. Figure B.7). When cross-referencing the WGBS data against the newly-assembled MN106-Ref genome annotation, high levels of DNA methylation (mostly mCG) co-localised with regions of dispersed repeats and TEs towards the centre of each chromosome. Conversely, methylation was depleted in gene-rich regions (Figure 9a,b). In line with this, DNA methylation was consistently high along TEs, particularly in the CG context (Figure 8c). In contrast to closely-related species *E. salsugineum* (Bewick et al. 2016; Niederhuth et al. 2016), DNA methylation dropped only slightly in regions flanking TEs, which might be related to the overall dense TE content in *T. arvense*.

In contrast to TE and promoter methylation, gene body methylation (gbM) is generally associated with medium-to-high gene expression levels (Xiaoyu Zhang et al. 2006; Zilberman et al. 2007). In *A. thaliana*, gbM occurs in ~30% of protein-coding genes in *A. thaliana*, with DNA methylation increasing towards the 3'-end of the gene (Xiaoyu Zhang et al. 2006), whereas the close relative to field pennycress, *E. salsugineum*, lacks gbM (Bewick et al. 2016; Niederhuth et al. 2016). gbM was also largely absent in *T. arvense* (Figure 8d), suggesting that it was lost at the base of this clade.

3.4 Discussion

Previous annotations obtained from the original assembly of *T. arvense* were herein further enriched with additional gene models for protein-coding loci in the newly-assembled genome, and now include non-coding genes for tRNAs, rRNAs, siRNAs, miRNAs, and snoRNAs, alongside predicted pseudogenes and TEs (Table 2). Identification of such features, alongside tissue-specific cytosine methylation and gene expression, aid in disentangling the coalescent and multifaceted nature of the epigenetic landscape and in understanding its role for example in regulating gene expression.

The improved genome assembly for line MN106-Ref, containing seven chromosome-level scaffolds, revealed two main features in its genomic landscape: a large repetitive fraction populated with TEs and pseudogenic loci in pericentromeric regions, and a gene complement densely concentrated towards the telomeres (Figure 9). Whilst the protein-coding gene fraction of the genome is similar in size to other closely-related Brassicaceae (Wang et al. 2011), the large repetitive fraction suggests an increased genome size driven by TE expansion (Beric et al. 2021). In addition, the spatial distribution of sRNA loci followed the gene density but was concentrated predominantly at the boundary between genes and TEs. Overall levels of methylation are high throughout the genome, likely owing to the large repetitive fraction where TEs are typically silenced in order to preserve genome stability, whereas gbM is low throughout the gene complement. This low gbM is not unsurprising relative to other Brassicaceae such as *E. salsugineum*, but differs from the model plant *A. thaliana* and the distribution typically expected in other plant species. Further exploration of mechanisms involved in the regulation of gbM may yield further insight to the evolution of epigenetic phenomena in *T. arvense* among other Brassicaceae. For example, the active demethylase REPRESSOR OF SILENCING 1 (ROS1) functions as genomic “methylstat” in *A. thaliana* (Lei et al. 2015; Zhang, Lang, and Zhu 2018). A helitron TE in the ROS1 gene promoter negatively controls ROS1 expression, whereas an RdDM target sequence in closer proximity to the 5' UTR upregulates transcription when methylated, thus leading to increased demethylase activity and concomitantly reducing genome-wide levels of methylation (including its own promoter). An exploration of the closest gene homolog in *T. arvense* reveals instead a copia TE occupying the region upstream in its promoter, indicating a possible functional difference which may have evolved independently of *A. thaliana*. A combined resource with different co-dependant feature annotations facilitates a layer of detail which is not appreciable e.g. with gene annotations alone.

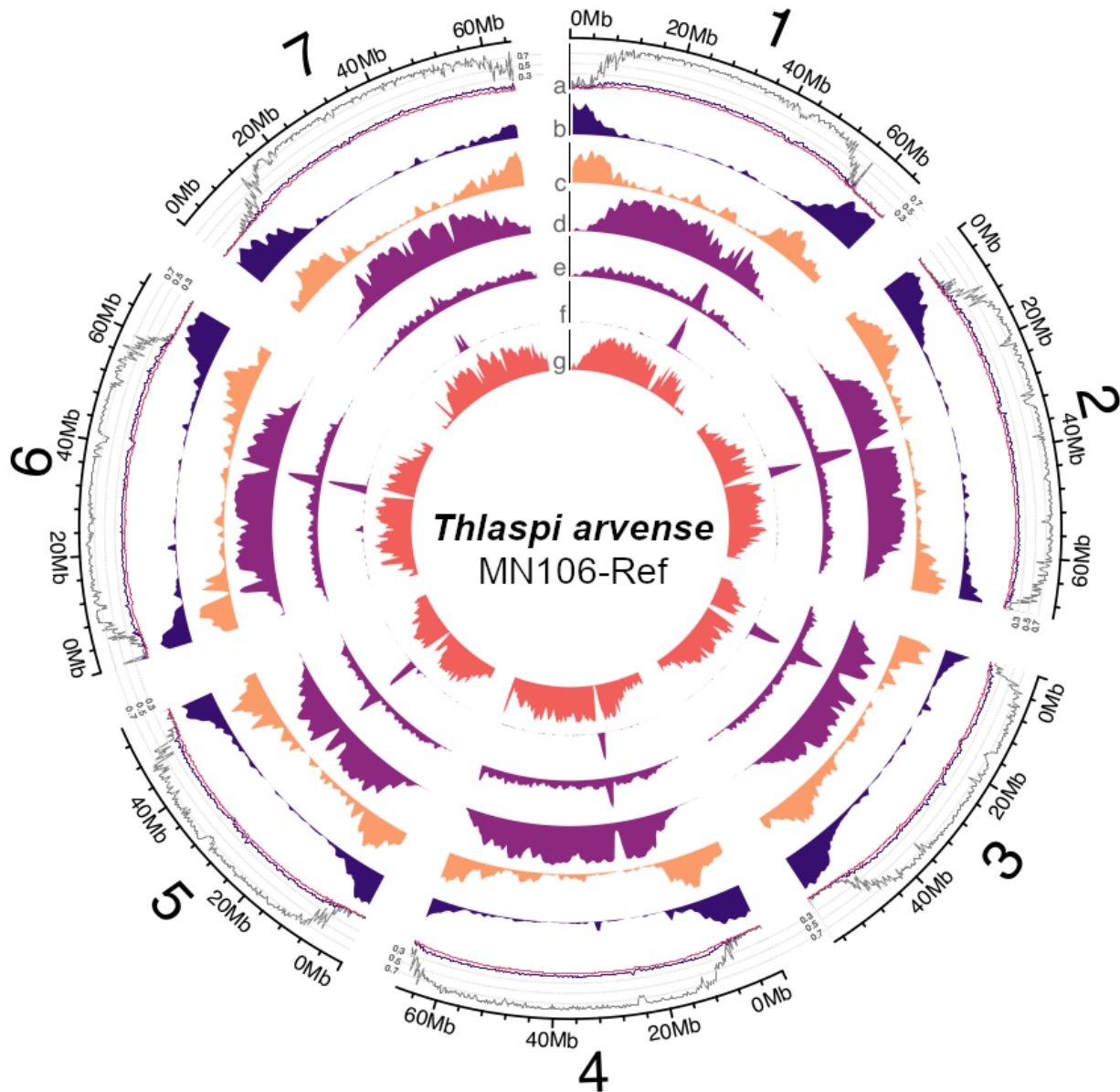


Figure 9. Overview of the seven largest scaffolds representing chromosomes in *T. arvense* var. MN106-Ref (*T_arvense_v2*). The tracks denote **a**) DNA methylation level in shoot tissue (CG: grey; CHG: black; CHH: pink; 200 Kbp window size), and density distributions (1 Mbp window size) of **b**) protein-coding loci, **c**) sRNA loci, **d**) Gypsy retrotransposons, **e**) Copia retrotransposons, **f**) LTR retrotransposons, and **g**) pseudogenes.

In addition, the newly improved assembly features will allow for efficient combining of traits and help accelerate future breeding practices in *T. arvense*, as it would provide knowledge about the gene localisation and the linkage of genes of interest. For example, the improved genome assembly has revealed that multiple domestication syndrome genes (ALKENYL HYDROXALKYL PRODUCING 2 - like, TRANSPARENT TESTA 8, EARLY FLOWERING 6) (Suppl. Figure B.8) are located on a single chromosome. With the availability of improved genomic resources, increasing interest has also turned towards understanding tissue-specific gene regulation to reduce

pleiotropic effects upon direct targeting of genes during crop improvement. The datasets generated herewith help elucidate the extent of tissue specificity and provide useful information for gene modification targets. For example, fatty-acid desaturase 2 gene (FAD2; Ta12495 - T_arvense_v1) is involved in the oil biosynthesis pathway and is expressed in many different tissues analysed in this study. FAD2 gene knockout should result in higher levels of oleic acid in the seed oil and provide an opportunity for pennycress oil to be used in food applications. It has been observed, however, that knockout mutants in pennycress display delayed growth and reduced seed yields in spring-types (Jarvis et al. 2021), and reduced winter survival in the winter-types (Chopra et al. 2019), as a purported consequence of its broad expression profile. Similarly, genes such as AOP2-LIKE (Tarvense_05380 - T_arvense_v2) have been targeted to reduce glucosinolates in pennycress seed meal for food and animal feed applications (Chopra et al. 2020). However, AOP2-LIKE, too, is expressed in many tissues during development, which might explain why knockout plants with reduced glucosinolate content are reportedly more susceptible to insect herbivores such as flea beetles feeding on rosette leaves and root tissues (Jez et al. 2021). The tissue-specific expression data suggest that, to overcome this challenge, one could alternatively target genes such as Glucosinolate Transporter 1 (GTR1; Tarvense_14683), which is expressed specifically in reproductive tissues. This might achieve the desired reductions of seed glucosinolates while avoiding developmental defects. Such approaches have been effectively used in *Arabidopsis* and many Brassica species (Andersen and Halkier 2014; Nour-Eldin et al. 2012).

In conclusion, the newly-assembled genome of *T. arvense* line MN106-Ref offers new insights into the genome structure of this species in particular, and of lineage II of the Brassicaceae more generally, and it provides new information and resources relevant for comparative genomic studies. The tools presented here provide a solid foundation for future studies both in an alternative model species to investigate epigenetics in plant ecology, and as an emerging crop.

4 Bisulfite Sequencing Methods

4.1 Introduction

Though it is by no means the only epigenetic mark prevalent throughout the genome, DNA methylation is involved in a wide range of molecular processes and is among the most-studied base modifications in this context. Chapter 1 of this thesis addresses the mechanisms of maintenance, distribution and potential ecological consequences of different patterns in DNA methylation variation- but how exactly do we detect these differences in the first place? One technique that has emerged at the forefront of epigenetic research is bisulfite sequencing: a distinct adaptation of short-read NGS technology which enables the characterisation of genome-wide methylation profiles at a nucleotide-level resolution.

The technique, devised by (Frommer et al. 1992) and refined for modern sequencing techniques by (Lister et al. 2008) and (Cokus et al. 2008), involves the treatment of extracted DNA from test samples with sodium bisulfite, a deaminating agent which mediates the conversion of unmethylated cytosine nucleotides into uracil. Cytosine bases that carry methyl groups (e.g. 5-methylcytosine, 5-hydroxymethylcytosine) are left unaffected by the treatment and remain in their original unconverted state. As the resulting single-stranded (ss)DNA then undergoes PCR amplification, uracil pairs with adenosine rather than the original guanosine during replication, which in turn pairs with thymine in the final, amplified product in place of the original cytosine. Bisulfite-treated samples can thus be subjected to standard sequencing protocols and used to generate sequencing reads, which carry epigenetic information. Once treated, the interpretation of sequenced reads effectively reframes the research question from a biological to a computational, algorithmic concern.

In line with typical NGS applications, the following step is usually read alignment of the sequencing reads to a reference genome assembly (as described in Chapter 1). Sequence alignment of short reads typically employs a “seed-and-extend” heuristic to shortlist possible locations on the reference genome which can be extended into a full alignment using a scoring approach. Such alignment presents some issues when handling bisulfite data, however, as thymine residues can no longer be considered as entirely independent entities to cytosine due to the base conversion during

treatment. Read alignment algorithms usually operate based on scoring matrices, from which an overall likelihood can be inferred for the alignment of two sequences based on the number and position of matches, mismatches, insertions, and deletions between nucleotides. With bisulfite-treated reads, the problem arises in that reference cytosines can conceptually now match with thymines, but not vice versa. Existing algorithms are often not built to handle this asymmetry between bases, so the solution is either to i) adapt these tools in some way further, or ii) to operate specifically with new algorithms designed for bisulfite data. Several tools now exist in representation of either category, including notably Bismark (Krueger and Andrews 2011) and BWA-meth (Pedersen et al. 2014), which adapt the popular standard aligners bowtie2 (Langmead and Salzberg 2012) and BWA (H. Li and Durbin 2009), and software such as segemehl (Otto, Stadler, and Hoffmann 2012, 2014) or ERNE-BS5 (Prezza et al. 2012) which are capable of interpreting bisulfite reads in their own right.

The principles of bisulfite sequencing notwithstanding, another important consideration when designing such an experiment involves the chosen strategy for library preparation. As with conventional NGS libraries, sequencing depth and coverage are also very important for bisulfite sequencing, in as much as the priority is to maximise the available information with which to address the study questions whilst balancing practical limitations, such as cost and time. For example, some studies may require investigation of genome-wide methylation patterns to assess overall variation or discover candidate markers, whereas others may prefer to focus on a reduced subset of the DNA in high resolution. Such approaches typically include Whole-Genome Bisulfite Sequencing (WGBS), Reduced-Representation Bisulfite Sequencing (RRBS), and further variations on these methods.

This chapter covers various technical concerns of bisulfite sequencing, from DNA extraction and library preparation to sequencing itself and the downstream extraction of cytosine methylation levels. The bioinformatic principles determine the data validity for answering the questions posed by the study, and an a priori consideration, therefore, is fundamental to the successful outcome of any such experiment. Finally, the specific limitations of bisulfite sequencing are discussed, and brief suggestions are given for alternative methods that might be used to address these issues.

4.2 Principles of Bisulfite Sequencing

When considering the epigenome, researchers may refer to changes in chromatin structure due to post-translational modification of histone proteins (Margueron and Reinberg 2010), populations of non-coding small RNA (ncRNA; sRNA) (Aufsatz et al. 2002; Cao et al. 2003; Matzke and Moshier 2014), chemical modifications to DNA sequences, or a combined effect of these factors. Most often, however, they are referring specifically to the methylome. That is to say: the distinct arrangement of methylcytosines present in the genome and the variation between different organisms, or tissues and cell types within species. This generalisation reflects both its epigenetic significance in a number of processes, and an overrepresentation of our current understanding of this DNA modification relative to other epigenetic mechanisms. It is just one phenomenon however within a multitude of epigenetic factors which often interact.

In plants, DNA methylation can affect both cytosine (Zhang, Lang, and Zhu 2018) and adenine (Ratel et al. 2006) nucleotides and has been associated with changes in gene expression (Jaenisch and Bird 2003; Xiaoyu Zhang et al. 2006; Lang et al. 2017), chromosome interactions (Grob, Schmid, and Grossniklaus 2014; S. Feng et al. 2014), and genome stability through the repression of transposable elements (Mirouze et al. 2009; Tsukahara et al. 2009; La et al. 2011). The modification can be further characterised as either 5-methylcytosine (5mC) or 5-hydroxymethylcytosine (5hmC) within the context of cytosine methylation. These subgroups may well have contrasting epigenetic functions, though in *A. thaliana* at least, no appreciable level of 5hmC has been observed in genomic DNA (Erdmann et al. 2014). The fraction of 5hmC present in RNA may be much higher (Huber et al. 2015), and indeed cytosine methylation is not a base modification that is limited to genomic DNA. The extent and prevalence of 5hmC in plants however is still debated among researchers (Mahmood and Dunwell 2019).

The underlying basis for the perceived emphasis on DNA cytosine methylation is due to the development of bisulfite sequencing as a means for studying epigenetics. Since its conception and initial application by (Frommer et al. 1992), the method has received much attention for its capacity to resolve DNA methylation patterns at the nucleotide level. This allows researchers to study the effect of differential methylation between organisms, tissues, or cell types on specific genomic elements such as gene bodies, promoter regions, or other regulatory motifs. This, in turn, provides

a roadmap for linking epigenetics to gene expression, heritability, or the activation of particular genes or transposons.

The basis for the method is the usage of sodium bisulfite. This chemical compound catalyses the hydrolytic deamination of cytosine to uracil via an intermediary sulfonation step from cytosine to unstable cytosinesulfonate (Hayatsu et al. 1970; Shapiro, Servis, and Welcher 1970). The loss of the amine group from cytosinesulfonate yields uracilsulfonate, which in turn is desulfonated under basic pH conditions to form uracil (Figure 10). Though in principle methylated cytosines can also react with sodium bisulfite, the presence of a methyl group on position five of the aromatic ring inhibits the process by an order of two magnitudes (Hayatsu and Shiragami 1979). This inhibition is sufficient to confer selectivity for the conversion of unmethylated cytosines. Unfortunately, this selectivity does not extend to a measurable differentiation between 5mC and 5hmC, which are therefore indistinguishable during the regular application of this method. During the sequencing

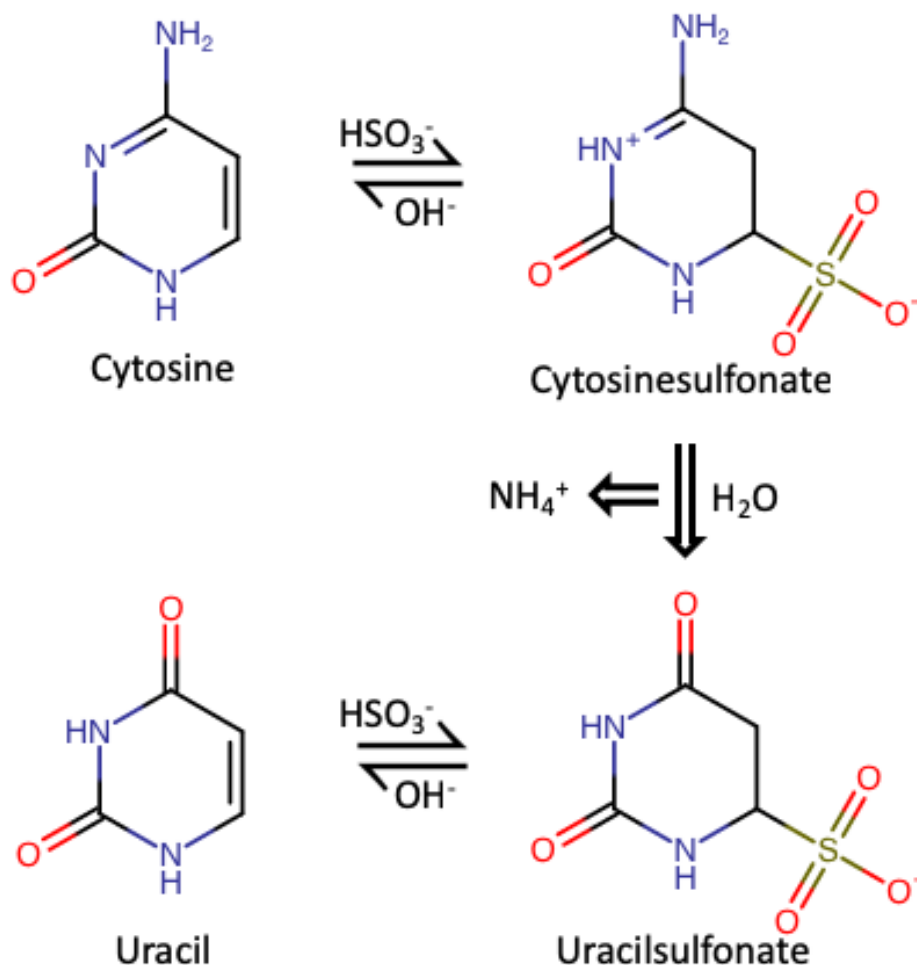


Figure 10. The reaction steps behind the conversion of unmethylated cytosine to uracil by sodium bisulfite.

4 Bisulfite Sequencing Methods

processes or preceding PCR reactions, uracil positions convert on the newly synthesised DNA fragment to thymine so that the sequencing reads contain the four DNA bases.

Once a given DNA sequence has undergone treatment with sodium bisulfite, any remaining cytosines present in the sequence can be inferred as methylated positions. The problem then progresses to the question of mapping these sequences back to their original location on the genome. As is often the case with NGS, it is very difficult to achieve a singular, continuous strand of DNA representing an entire chromosome. DNA stability is such that the molecules are easily fragmented during library preparation, and the sequencing approach itself is often restrictive in terms of the length of DNA that can be sequenced in a single iteration. The fragmentation is further confounded by the harsh bisulfite treatment, which undermines long-read sequencing technologies such as PacBio (Eid et al. 2009; Uemura et al. 2010) or Nanopore (Cherf et al. 2012; Mikheyev and Tin 2014) and reduces the cost-benefit ratio for using them.

The dominating circumstance is one where the sequencing data consists of many, short bisulfite-converted DNA sequences (typically Illumina) that subsequently need to be aligned to a reference genome to determine the relationship of methylated positions with nearby genomic elements. In this regard, it is imperative that a high-quality reference genome is available for the species of interest. Any improvements made to the existing genome assembly contribute towards mapping precision and reducing any false correlations between methylation patterns and nearby loci.

It is not a simple task to map bisulfite-treated reads to the reference genome as it would be performed with conventional NGS data. The challenge arises from the many unmethylated cytosines, which have since been converted to thymines. The fraction of methylated cytosines varies much between different plant species (Feng et al. 2010; Zemach et al. 2010; Niederhuth et al. 2016), but for example with 5.26% of genomic cytosines reportedly methylated in *A. thaliana* (Lister et al. 2008), given a bisulfite conversion rate of 99.14% and a GC content of ~36%, then theoretically ~17% of the total genome space on each strand would be artificially replaced with thymines, where before they were cytosines. In a standard read alignment procedure, this would result in many differences in the alignment of reads to the reference sequence due to the high fraction of C>T mismatches occurring in place of matches. Specifying that all cytosines and thymines should match symmetrically might provide a solution but would result in many spurious alignments arising from reduced base complexity. Asymmetrical base-matching would result in greater precision, as after bisulfite treatment only thymines present in the sequencing reads can

potentially match cytosines in the reference genome. It does not follow conceptually that thymines in the reference genome might be able to match cytosines in the reads, as the reference genome is not treated.

Standard alignment tools are often not designed with this base-matching asymmetry in mind, however, so the necessary solution is to make further adaptations to these tools or operate specifically with new algorithms designed to handle bisulfite data. As mentioned previously, several specific alignment programs for bisulfite data exist. Prominent among them are Bismark (Krueger and Andrews 2011) and BWA-meth (Pedersen et al. 2014), relying on popular alignment algorithms for conventional data, or Segemehl (Otto, Stadler, and Hoffmann 2012, 2014) and ERNE-BS5 (Prezza et al. 2012), with more specific algorithms. Given the number of tools available, it is important to understand how they perform under different experimental conditions and circumstances. For this purpose, a number of benchmark studies have been undertaken which focus for example on algorithmic differences (Tran et al. 2014), combinations of pre- and post-processing techniques (Tsuji and Weng 2016) or just a small range of tools on model data (e.g. human) (Chatterjee et al. 2012; Kunde-Ramamoorthy et al. 2014). As of yet however there is no comparable analysis of alignment performance in emerging applications such as with non-model plant data, which often presents its own challenges based on the underlying quality and complexity of non-model reference sequences. This gap is further addressed by a new benchmark analysis presented in Chapter 5 of this thesis.

Another problem that arises during bisulfite sequencing is the loss of variant information following the sodium bisulfite treatment. Under a standard NGS experiment, single nucleotide polymorphisms (SNPs) are identified where a number of reads overlapping a single nucleotide position on the reference genome might indicate a deviation from the reference base in that position. A model of genotyping of SNPs is necessary for example in identification of which samples belong to different variants or strains. This may be necessary also during epigenetic studies, as it is often of crucial importance that genetic variation is kept to a minimum to reduce confounding genetic effects. Any true SNPs in CT context are however obscured by the artificially-converted bases arising due to bisulfite treatment. It is possible to retrieve this variant information by comparing the converted bases to their complementary bases on the opposite strand, as only true SNPs should have the correct sequence complement. This paradigm is utilised by a number of bisulfite-aware variant callers, such as BISCUIT (<https://github.com/huishenlab/biscuit>), Bis-SNP (Yaping Liu et al. 2012), BS-SNPer (Gao et al. 2015), gemBS (Merkel et al. 2019), and

MethylExtract (Barturen et al. 2013), however the tools available for this purpose are both sparse, and fraught with a number of technical shortcomings, and consequently this task is yet to be implemented efficiently in comparison to tools developed for conventional sequencing data. This gap is further addressed by a new method presented in Chapter 6 of this thesis.

Finally, following a successful read alignment, the task of obtaining methylation levels over each cytosine is typically performed in a similar manner to variant calling, but without a complex model for genotyping. Instead, each cytosine position is extracted from the alignment as before and the methylation level determined by majority voting. The ratio of reads with either cytosine or thymine in that position is used to calculate an overall methylation percentage or rate. The total collection of positions can be further subset into different genomic sequence contexts such as CG, CHG, or CHH, where H can be either A, T, or G (see Chapter 1 section 1.2.2). This subsetting later allows for downstream analyses of methylation patterns.

4.3 Experimental Design

As with many ecological studies involving next-generation sequencing, the experimental design is often based heavily around one fundamental trade-off, which is paramount to achieving a level of statistical power appropriate for the aim of the study. The trade-off refers to the balance of maximum sequencing coverage relative to the practical limitations of the study, such as cost and time. Next-generation sequencing is expensive, with costs driven by the quantity of material to be sequenced. This can be delineated to the total genome size, number of replicates, level of sequence coverage, and sequencing technology itself. Therefore, an ideal study seeks to define and optimise these factors a priori to carrying out the experiment and the subsequent downstream analyses.

Next-generation sequencing is a fast-developing field, with several commercial technologies currently being available to address various experimental needs and applications. These can broadly be categorised into short-read technologies (~50-900 bp), which tend to be cheaper with higher per-base accuracy and throughput than the counterpart, long-read technologies (~1-500 kbp) (Goodwin, McPherson, and McCombie 2016; Li and Harkess 2018). The general advantage of obtaining long reads is that they are less prone to assembly and mapping-related errors, making them suitable for resolving true genome arrangements even in repetitive regions and regions of low complexity, e.g. found in heterochromatin regions. Due to the issues mentioned above with generating long fragments from sodium bisulfite-treated DNA samples, however, these advantages

are frequently lost. Short-read sequencing technologies, therefore, have been predominantly selected for methylation analyses.

Among these short-read sequencing technologies, Ion Torrent (Rothberg et al. 2011), Illumina (Bentley et al. 2008), and SOLiD (Shendure et al. 2005; McKernan et al. 2009) have all been used successfully for bisulfite sequencing (Cokus et al. 2008; Lister et al. 2008; Lang et al. 2017; Mirouze et al. 2009; Bormann Chung et al. 2010; Pabinger et al. 2016; Venney, Johansson, and Heath 2016). Each may be appropriate depending on the desired read size and number of reads per run, which is influenced by the size of the genome and the nature of the study in question (i.e., sample size). The most extensively used of these is Illumina, wherein the HiSeq 2500, 3000, 4000, or HiSeqX and NovoSeq (for larger projects) are all appropriate for high-throughput applications. A reasonable comparison between these and some alternative systems is given by (Grehl et al. 2018).

A problem arises for bisulfite sequencing from the requirement of many sequencing machines that nucleotides should be present in the DNA fragments in roughly even proportions to perform efficient base calling. Following the bisulfite treatment, cytosines are underrepresented. To still enable high-throughput sequencing, additional DNA with balanced base proportions and known sequence (e.g. PhiX) is added or "spiked-in" during library preparation. This DNA standard is sequenced together with the target DNA and later filtered out. The disadvantage is that it reduces the potential sequencing coverage considerably because the machines may require 20-40% spiked-in standard DNA. The HiSeq 2500 is notable here as it benefits specifically from an optimisation of the cluster calling algorithm, which allows for the handling of bisulfite-treated libraries without the need for additional base proportion balancing (Grehl et al. 2018). Thus, the HiSeq 2500 is a good baseline to act as a starting point for designing bisulfite sequencing experiments.

As methylation calling is calculated from the number of overlapping reads aligning to each position, it is clear that the statistical power increases with greater sequencing depth. In an ideal scenario, the sequencing depth would be uniform and consistent across all positions of interest to the study, but this is rarely the case. Sequence-related biases in the random DNA fragmentation (Poptsova et al. 2014) and PCR amplification (Kozarewa et al. 2009; Aird et al. 2011) stages of library preparation impede uniformity and lead to coverage underrepresentation in regions of extreme GC-content (Benjamini and Speed 2012). A straightforward approach in mitigating these issues is to select a level of coverage that ensures a minimum lower bound in the majority of regions where the distributed coverage is lower than the mean. Nevertheless, the bisulfite treatment itself can yet

4 Bisulfite Sequencing Methods

induce new biases, or exacerbate existing ones, for example in terms of cytosine depletion, PCR amplification, cytosine modification, and conversion artefacts; each of which have varying consequences for downstream methylation analysis (Olova et al. 2018).

To give a point of reference, (Ziller et al. 2014) found that coverage between 5-15x was optimal in terms of statistical power for detecting differentially methylated regions between a range of human tissue and cell types in the CG context. Beyond that, resources would be better allocated towards expanding the number of biological replicates, starting at a minimum of two to capture within-group variance. In plants, it would be wise to consider this point of reference as a bare minimum for a homozygous diploid. Both the heterozygosity and the ploidy level undoubtedly influence the minimum level of coverage due to the increased variation on single positions, which may lead to greater within-group heterogeneity necessitating a larger number of replicates. The highly-repetitive regions, as well as low-complexity regions often found in plant genomes, are also notoriously difficult to map to, leading to many multi-mapped reads and alignment ambiguities (Treangen and Salzberg 2011). In the absence of long-read data, this problem can be mitigated by increasing the coverage and using paired-end (PE) sequencing. Furthermore, the magnitude of methylation differences are usually less pronounced in the CHG and CHH contexts, in comparison to the more “binary” CG context (Cokus et al. 2008; Lister et al. 2008), thus requiring more power for detection. If the study seeks to capture differences in these contexts then an increase in both sequencing depth and the number of replicates should be considered relative to CG alone.

Once the optimal level of coverage and number of replicates have been decided, it may be the case that the total genome size for the species of interest pushes the cost outside the range of affordability. In these instances, it is sometimes possible during library preparation to subset the material and exclude regions of minor or no relevance to the scope of the study. Whether this is appropriate or not depends on whether the scope of the research question requires whole genome data. An overview of both whole genome and reduced representation methods are given in the following section.

4.4 Library Preparation

4.4.1 Whole Genome Bisulfite Sequencing (WGBS)

WGBS is the practice of applying bisulfite sequencing on a genome-wide scale, capturing all regions, and attempting to define global methylation patterns in each sample. This method is

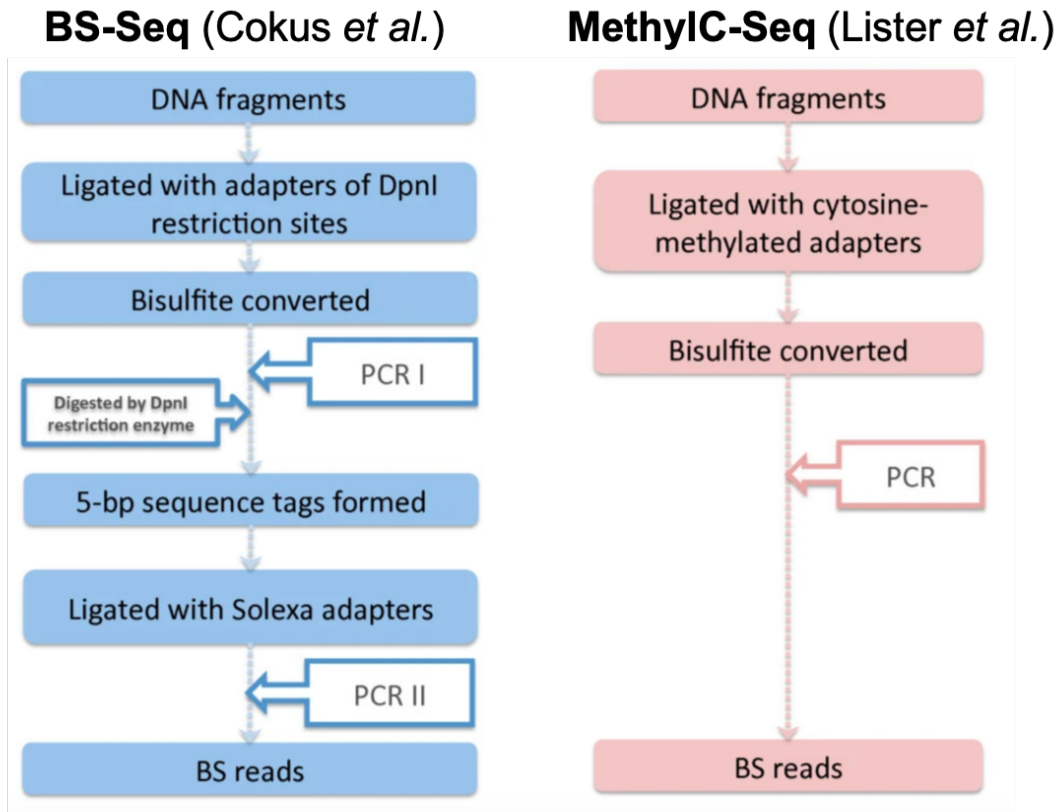


Figure 11. An overview of BS-Seq (Cokus *et al.* 2008) and MethylC-Seq (Lister *et al.* 2008) library preparation protocols for bisulfite sequencing. Adapted from Chen *et al.* (2010) “BS Seeker: precise mapping for bisulfite sequencing”.

appropriate when the study question is broad in scope or if prior information on the genomic regions of interest is limited. It can be considered the "go-to" approach when other methods for concentrating the sequencing on reduced subsets of the genome are either unavailable or inappropriate for the study in question.

There are two main variations of WGBS library preparation, known as BS-Seq (Cokus *et al.* 2008) and MethylC-Seq (Lister *et al.* 2008, 2009) (Figure 11). In terms of the protocol, they differ primarily in the number of PCR steps and when the ligation of sequencing adapters occurs relative to the treatment with sodium bisulfite. Many sequencing technologies require specific sequencing adapters to facilitate base calling on selected DNA fragments. In the case of Illumina, these adapters are bound to complementary sequences on the flow cell, forming clusters to be sequenced-by-synthesis. If the adapter is not present, the DNA molecule is simply washed off the cell, and no information is retrieved. The issue here is that the bisulfite treatment alters the sequence of these adapters wherever there are unmethylated cytosines present, rendering them incompatible with the complementary sequences on the flow cell. MethylC-Seq addresses this by using custom, fully methylated adapters that remain unaffected by sodium bisulfite. In contrast,

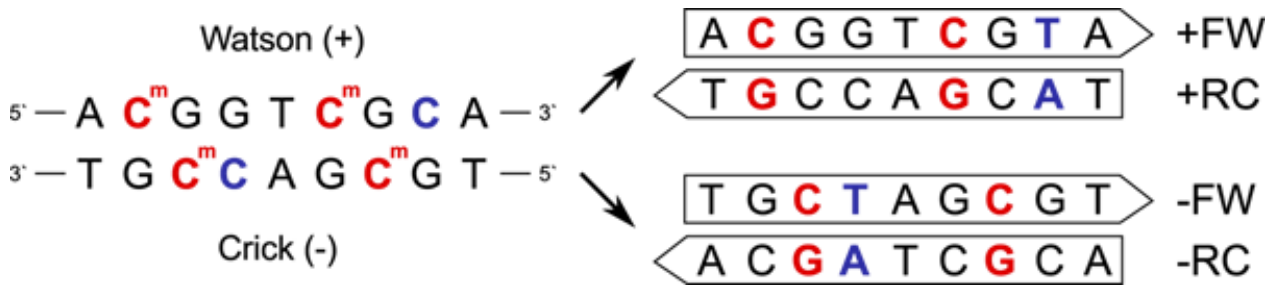


Figure 12. Strand-specific point mutations in the newly-synthesised strand resulting from bisulfite conversion of unmethylated cytosines.

BS-Seq circumvents the issue by ligating the adapters only after the bisulfite treatment has taken place.

In principle, the approach of BS-Seq seems more straightforward. However, ligating the sequencing adapters after the bisulfite treatment presents another problem: the two strands of DNA are no longer complementary to each other and hence remain in a single-stranded state. Typical sequencing adapter ligation requires duplex DNA, therefore an additional round of PCR is necessary before adapter ligation can occur. This PCR step begins with both the bisulfite-treated Watson (+FW) and Crick (-FW) strands of the original DNA (Figure 12), and generates reverse complementary strands (+RC and -RC, respectively). The result is a set of four distinct sequences which are indistinguishable from each other by the sequencer. Strand-specificity is therefore lost, and additional bioinformatic processing is required to resolve which reads belong to which strand. In MethylC-Seq, only the +FW and -FW sequences are present, and strand-specificity is thus cleanly preserved during sequencing. It becomes more complex with paired-end data however, as the +RC and -RC strands are present as well following sequencing of mate 2.

A more recent variation of these approaches has also been developed, known as post-bisulfite adapter tagging (PBAT) (Miura et al. 2012). In this case, the bisulfite conversion process itself is first used to fragment the genomic DNA. Adapter ligation is then facilitated by two rounds of random priming extension in place of PCR, thereby maintaining strand-specificity while avoiding any denaturation of adapter-ligated DNA. The real advantage of this method, however, is its sensitivity in handling sub-microgram quantities of DNA without the need for additional amplification, contrary to MethylC-Seq where the bisulfite treatment often fragments adapter-ligated DNA templates which then cannot be used during sequencing. In such a case, the remaining DNA may need to be amplified to achieve a reasonable DNA mass for sequencing, but this

amplification risks inducing PCR artefacts. The approach of PBAT can circumvent the need for PCR amplification on sub microgram quantities of DNA. Still, it should be noted that random primer extension is subject to its own biases. Sequence-specific site preferences can give rise to "pile-ups" of reads, and differential priming between methylated and unmethylated alleles has been hypothesised. Therefore, it may be preferable to run MethylC-Seq with a very low number of PCR amplification cycles (e.g., ~4) in cases where sample availability is not strictly limited. Yet more recent approaches such as TET-assisted pyridine borane sequencing (Yibin Liu et al. 2019) even attempt to circumvent bisulfite treatment altogether, reportedly allowing for higher mapping quality and non-fragmented duplex DNA after the conversion of methylated cytosines into thymines by alternative chemical means. Such methods may avoid some fragmentation issues of harsh bisulfite treatment, but commercial kits for library preparation have only been introduced recently by comparison to previous methods, and the application thus far in the literature is scarce.

Regardless of the type of bisulfite library selected, at least two cycles of post-bisulfite PCR are necessary to facilitate the conversion of uracil to thymine before sequencing can occur. For these PCRs, the presence of uracil in the sequence precludes the use of many standard, high-fidelity polymerase enzymes with proofreading mechanisms such as Phusion (Thermo Scientific) or KAPA HiFi (Roche). On encountering uracil, these enzymes stall as they await base excision repair (Lindahl and Wood 1999; Greagg et al. 1999). Fortunately, there are alternatives available, such as PfuTurbo Cx (Agilent) or KAPA HiFi uracil+ (Roche), which are more specifically suited for bisulfite sequencing.

Several commercial kits are readily available for carrying out bisulfite conversion itself. Depending on the sample DNA quantity and library preparation methodology, the aim is to achieve maximum conversion efficiency relative to optimal DNA recovery. High temperature, high bisulfite molarity, and long incubation times are more likely to yield complete bisulfite conversion but degrade much of the DNA in the process. Incomplete conversion, however, leads to an overestimation of methylation levels on unconverted cytosines. With this trade-off in mind, a good evaluation of modern kits was provided by Kint et al. (2018), where EpiTect Bisulfite (Qiagen), EZ DNA Methylation-Gold (Zymo Research), and EZ DNA Methylation-Lightning (Zymo Research) kits were each cited for high performance with regards to several study-dependent factors. Conversion efficiency within bisulfite-treated samples is then typically estimated through the use of control sequences, consisting of a known quantity of unmethylated DNA within the sample. Historically, the conversion rate has also been estimated from non-CG cytosines in mammals (Hodges et al.

4 Bisulfite Sequencing Methods

2009), which is inappropriate for plants where DNA methylation occurs also in CHG and CHH contexts. Alternatively, the mitochondria or chloroplast genomes have been used, as both organelles have been widely observed to escape DNA methylation (Marano and Carrillo 1991; Vanyushin and Kirnos 1988). This may not be entirely reliable either, however, as there is some conflicting evidence of DNA methylation which has been reported in both (Šimková 1998; Fojtová, Kovarík, and Matyásek 2001). Furthermore, the situation is often exacerbated by incomplete plastid genome assemblies (particularly in non-model species), or by the presence of nuclear-inserted plastid DNA (Michalovova, Vyskot, and Kejnovsky 2013). The most reliable method in plants is therefore to use a "spike-in" of DNA from another source. The enterobacteria phage Lambda (~0.1% w/w) is often used, which is shown to be virtually devoid of 5mC when propagated on mutant bacteria strains lacking DNA methylase activity (Hattman, Schlagman, and Cousens 1973). Reads aligning to the Lambda genome can then indicate the level of bisulfite conversion, as in theory, all cytosines should have been replaced with thymines.

In addition to Lambda, the bacteriophage PhiX is commonly used as a "spike-in" to balance base proportions (Raine, Liljedahl, and Nordlund 2018). During the initial cycles of Illumina sequencing, the phasing/pre-phasing, colour matrix corrections, and pass filter calculations are influenced by the flow cell imaging. In bisulfite-treated DNA, there is a notable deficiency in cytosine bases and the fluorescent colour associated with it, which can skew the base-calling algorithm during this normalisation process. Adding PhiX (Sanger et al. 1977) or any other well-balanced DNA to the sequencing library allows the Illumina sequencing to proceed unaffected. Another possibility is to multiplex both a bisulfite-treated library and a conventional, untreated library. This way, spiking can be omitted, and cytosine methylation and single nucleotide polymorphisms (SNPs) can be obtained from a single sequencing run.

Once preparation is complete, it is standard practice to perform library quantification and normalisation, using, for example, Qubit / PicoGreen assay or qPCR measurement. It should be noted during this step that methods that estimate only the total quantity of DNA may fail to give an accurate representation of the adapter-ligated DNA, particularly in MethylC-Seq libraries due to the aforementioned fragmentation caused by the bisulfite treatment. For this reason, it is typically recommended to use a BioAnalyzer for sizing only and qPCR to quantify the final library for bisulfite sequencing.

4.4.2 Reduced Representation Bisulfite Sequencing (RRBS)

RRBS is similar to WGBS in many ways, but differs primarily by adding an initial selection procedure at the beginning of the library preparation. It was developed by (Meissner et al. 2005) to generate large-scale sequencing data, with a lower resolution than WGBS, which evenly represents the genome, though with the option to focus either on eu- or heterochromatin. This reduces the sequencing cost compared to WGBS but results in the loss of much sequence content that could otherwise be relevant to the biological interpretation. Depending on the restriction enzyme used, the enriched fraction is typically expected to be less than ~1% of the whole genome size (Meissner et al. 2005).

Sample DNA is first subjected to a restriction endonuclease that targets a specific sequence context depending on the local cytosine methylation status. The sequence flanking the recognition site is then sequenced, providing information on the methylation status of many cytosines adjacent to the recognition site, which can principally be in all three sequence contexts. A typical enzyme used is MspI, which targets CG sites in the specific sequence 5'-CCGG-3'. MspI cannot cleave when this specific recognition sequence is symmetrically methylated, and thus focuses on weakly methylated euchromatin rather than the heavily methylated heterochromatin in the chromosomes. Different sequence contexts require different enzymes, although this application has not been broadly applied in non-CG contexts. The enzymatic digestion produces fragments that can be size selected, usually following some additional end-repair and A-tailing depending on which enzyme was used. The rest of the library preparation follows closely with that which was outlined previously for WGBS and unfortunately suffers from the same loss of strand-specificity as BS-Seq.

4.4.3 Target capture bisulfite sequencing

RRBS is a beneficial technique when the aim is to sequence many biological samples, for example, to study population genetics or when the studied organism has a very large genome, as in many coniferous trees, for example (De La Torre et al. 2014). The technique further allows to roughly direct the analysis either to heterochromatin or euchromatin or, depending on the genome in question, enrich promoter or gene-body sites by choosing the appropriate cleavage enzyme. However, besides this possibility of setting a rough focus of the study, the idea is to provide a valid representation of the genome through a sample of random sequence reads scattered across the genome. Though, it may be desirable in a project to set the target more specifically to a particular region in the genome. This can be achieved conversely through “target capture”, which can be applied before or after bisulfite conversion (Wreczycka et al. 2017). Different techniques usually

involving the hybridisation of genomic DNA with the complementary of a known piece of the target sequence, combined with bisulfite conversion and followed by the above-described processing of converted DNA, enable the inference of the methylation status of a specific target location in the genome of interest. Such techniques may be helpful when, for example, describing the methylation status of a known promoter region is the aim of the investigation, and genome-wide data is thus superfluous.

4.5 Bioinformatic analysis of bisulfite data

Once sequencing has taken place, the question of identifying DNA methylation is effectively reframed to a computational concern. As with standard NGS, the basic workflow tends to involve initial quality control (QC) of the generated raw reads, followed by mapping these reads to a reference genome, to produce alignment files that are the basis of downstream analyses. Methylated positions can then be extracted from these files in a much similar manner to variant calling. Bisulfite sequencing, however, presents its own challenges, particularly during the read alignment step. This section explores common issues and significant divergences from standard practices in bioinformatics.

4.5.1 Quality Control

Like all reads generated via sequencing by synthesis, bisulfite reads are subject to a drop in quality towards the 3'-end due to the propensity of base-calling errors to accumulate following failures in the synthesis process. During a single cycle of Illumina sequencing, the next base in the read is incorporated into the template strand together with a reversible terminator molecule containing a fluorescent tag (see Chapter 1 section 1.3.1). The terminator prevents the next base in the sequence from being incorporated, so the sequencer can appropriately read the colour of the fluorescent tag and identify the current base. The molecule is then cleaved to facilitate the next cycle, thus repeating the process for the next base in the sequence. During a single cycle, the terminator molecule can be either not cleaved or cleaved too early, resulting in the incorporation instead of two bases. If such an error occurs, the strand becomes "out of phase" compared to the other strands of the cluster, for all remaining cycles, which makes it more difficult for the imaging system to assess the correct base colour within the sequence cluster. The consequence is a quality drop for the complete cluster. This phenomenon is known as phase-shifting and is usually corrected by trimming some bases that fall below a quality threshold (e.g. phred score < 20) from the 3'-end of a read, limiting the negative effect on the previous correctly sequenced bases.

Another commonly-encountered issue is the tendency to sequence into the adapter sequence on DNA fragments smaller than the total number of Illumina cycles. These sequence subsets are not part of the original DNA, making it much more difficult to map such reads to their true location on the reference genome. Fortunately, the adapters are synthetically designed, distinct sequences which are therefore known and can thus be readily identified. By considering the overlap of the known adapter sequence at the 3'-end on each read, the reads can again be trimmed to remove the DNA that is not part of the original sequence.

These common problems can be identified with a standard QC tool such as FastQC (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>), and frequently occur in standard sequencing and bisulfite sequencing data. As such tools are usually not designed for bisulfite sequencing, they may also flag errors such as unbalanced base proportions with as high as ~50% thymine content in read one (or adenine content in read two). If dealing with RRBS data, which has been digested for example with MspI, there is also the possibility that non-random sequence content is flagged at the 5'-ends of reads, as digested fragments always start with a C base. So long as there is confidence that the standard precautions were taken during library preparation, these warnings can be safely ignored at this stage.

Conversely, the other facet of using standard QC tools that are not designed for bisulfite sequencing is the tendency to miss bisulfite-related sequencing problems. One such problem can occur during the initial DNA fragmentation step of the library preparation procedure, which often leaves protruding 5' and 3'-ends that must be restored to double-stranded DNA by a process known as overhang end-repair (Poptsova et al. 2014). The incorporation of unmethylated cytosines during this step can introduce artificially low methylation rates at each end of the DNA fragment, which cannot be detected in standard QC (Lin et al. 2013). Another such issue is thought to occur due to the re-annealing of single-strand sequences adjacent to the methylated sequencing adaptors during MethylC-Seq, which partially restores double-strandedness thereby affording a measure of protection from the bisulfite treatment (Lin et al. 2013). Therefore, there is a tendency for bisulfite conversion failure to be enriched towards the 5'-end of reads, leading to artificially high methylation rates. BS-Seq and PBAT libraries should theoretically avoid this bias due to the adapter-ligation occurring after the bisulfite treatment.

As neither of these issues causes changes to the DNA sequence, they can only be detected once methylation calling has occurred (after read alignment). The standard procedure is to look at the total distribution of methylation levels across the average length of the reads in an approach known as M-bias analysis (Hansen, Langmead, and Irizarry 2012; Lin et al. 2013). A uniform distribution is expected across the read length, but spikes in methylation level can be observed at each end of the distribution. As the sequence information at each end remains unaffected, they should be used for the alignment and not be clipped similarly to quality trimming for repeated read alignment. Instead, the start and end positions of reads are "masked" from a follow-up repeat of the methylation calling procedure, depending on the deviation of their methylation status from the uniform distribution. It should be noted that a significant variation of read lengths reduces the accuracy of this step.

4.5.2 Read Alignment

When all quality concerns are resolved, the next step in the basic workflow is to align these reads to an appropriate high-quality reference genome. If a reference genome is not available, *de novo* assembly is first required before any methylation information can be retrieved. The availability of a good reference genome is fundamental to DNA methylation analysis, when the project aims to relate methylated positions to nearby annotations such as gene bodies, promoter regions, or transposable elements. The higher the genome quality and relatedness of the genome to the test sample, the more confidence in the findings of the study (Mardis et al. 2002). Unfortunately, this degree of confidence is not something that can be quantified easily and directly. Therefore every effort should be made to ensure the qualitative validity of the reference genome prior to analysis.

With standard NGS data, mapping typically involves the use of dynamic programming to determine the best alignment for a given read according to a scoring matrix (see Chapter 1 section 1.3.3). Positive scores are given for base matches, or certain types of mismatches, whereas penalties are given for other mismatches and positions where insertions or deletions (indels) are present. The cumulative score is then compared to other potential alignments above a set threshold, and in most cases, only the best one is selected as the most likely point of origin for the read. Mapping bisulfite-treated DNA however presents a challenge. The majority of cytosine positions on the reference genome are likely unmethylated in the test sample (Wagner and Capesius 1981; Leutwiler, Hough-Evans, and Meyerowitz 1984) and therefore represented as thymines in the reads following bisulfite conversion. Aligning these reads results in many CT mismatches, which negatively influence the scoring matrix and significantly inhibit successful read mapping (Figure 13). It is not enough to

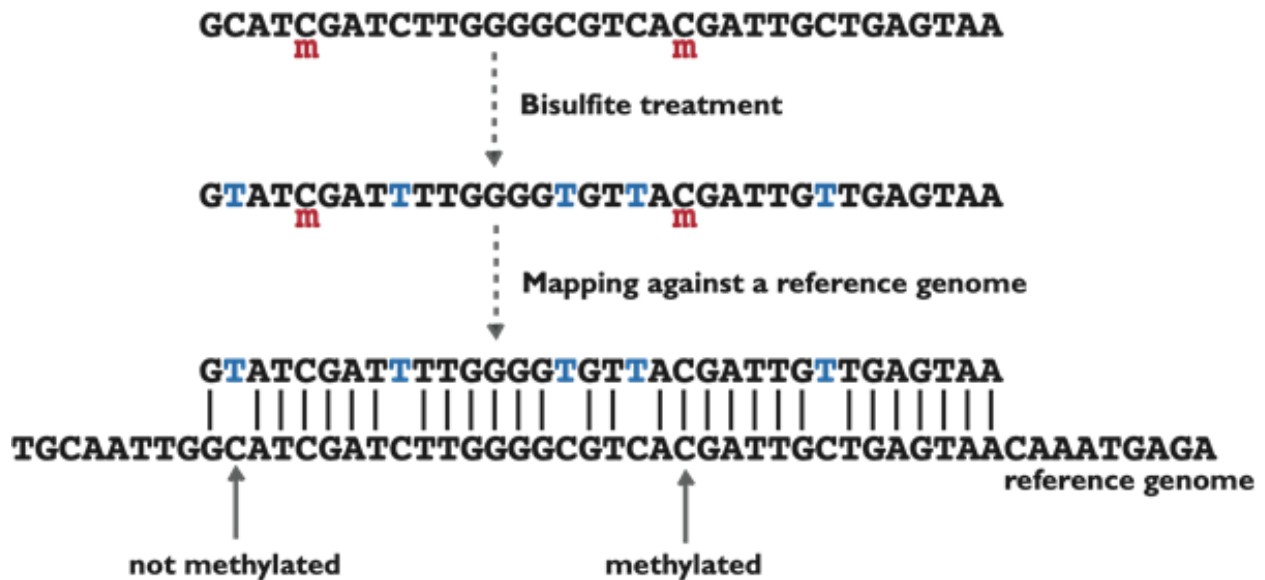


Figure 13. Alignment mismatches with the reference genome, as a result of bisulfite conversion of unmethylated cytosines.

simply allow for a higher number of errors, as this would only obscure the correct alignments through an increased number of false positives. To further complicate the issue, bisulfite conversion of DNA fragments results in two strands that are no longer complementary to each other. In single-end BS-seq or paired-end MethylC-seq, this means that four distinct sequences are now present, each one varying to some degree from the original DNA (Figure 12). From the original, untreated Watson (+) strand, the first mate pair is the direct bisulfite-converted variant (+FW), and the second is the reverse complement of this (+RC). From the original, untreated Crick (-) strand, the first mate pair is the direct bisulfite-converted variant (-FW), and the second is the reverse complement of this (-RC). In standard NGS, it is simply the case that the second mate-pair obtained from one strand aligns to the other strand, but in bisulfite sequencing, this no longer holds. In BS-Seq libraries, this is compounded further because the method already encompasses all four-strand variants, even in single-end sequencing. In this case, the directionality relative to the strand (indicated by the box arrows in Figure 12) is thus lost, which is why it is sometimes referred to as an unstranded bisulfite sequencing protocol.

One potential solution for this alignment problem is to adjust the scoring matrix so that a mismatch of thymine to cytosine is instead treated as a match. This can be implemented in standard sequence aligners by "collapsing" the genetic alphabet in both the read and the reference genome so that all cytosines are rewritten as thymines (Figure 14A). Mapping is then performed normally, and the methylated positions are retrieved through post-processing based on the composition of the pre-

A) Collapsed alphabet



B) Asymmetric matching



Figure 14. Consequences of collapsed alphabet versus asymmetric approaches during the seed-and-extend alignment procedure for bisulfite sequencing data.

collapsed sequences. This procedure results in two undesirable scenarios that do not fit conceptually: 1) true thymines in the read match with cytosines in the reference genome, and 2) cytosines from the read (indicating methylated positions) match thymines in the reference genome. Such a solution undoubtedly produces many false positives and obscures the correct read alignments. Bisulfite read aligners such as Bismark (Krueger and Andrews 2011), BSmooth (Hansen, Langmead, and Irizarry 2012) (in bowtie2 mode), BS-Seeker (Chen, Cokus, and Pellegrini 2010; Guo et al. 2013), and BWA-meth (Pedersen et al. 2014) follow this strategy.

A better strategy would be to allow matches between thymines and cytosines, but only between read-based thymines and reference-based cytosines (not vice versa). In this case, it is still possible (and unavoidable) for true read-based thymines to match incorrectly with reference-based cytosines, but methylated cytosines are correctly considered mismatched with thymines in the reference genome thus reducing false positives (Figure 14B). This asymmetric base scoring is not easy to implement in most index structures (e.g. Burrows-Wheeler transform, suffix arrays) used in standard sequence aligners. Therefore, specialised read alignment software is required that is explicitly designed for bisulfite sequencing. Such tools include BSMAP (Xi and Li 2009), BSmooth

(Hansen, Langmead, and Irizarry 2012) (in merman mode), and ERNE-BS5 (Prezza et al. 2012). The specialised read aligner segemehl (Otto, Stadler, and Hoffmann 2012, 2014) uses collapsing (Figure 14A) during a starting step before switching to asymmetric matching (Figure 14B) in the following step. A common drawback of these methods is the increased memory consumption and processing time relative to tools that rely on a collapsed alphabet.

Whichever approach is followed, the entire process likely has to be repeated to account for both CT and GA conversions due to the aforementioned loss of complementarity between a given sequence complement and the opposite strand (Figure 12). The possibility of four distinct sequences in bisulfite sequencing, as opposed to two in standard NGS, dictates that two distinct variants of the reference genome are required to resolve the best alignments. These two alignment procedures may either be run in parallel, as is the case in Bismark (Krueger and Andrews 2011), or consecutively as is the case in segemehl (Otto, Stadler, and Hoffmann 2012, 2014). Either way, a process of post-filtering and comparison must be made to unify the resulting alignments.

One problem that might arise during sequencing is the potential for genomic rearrangement (Saxena, Edwards, and Varshney 2014) that may have occurred in the test sample relative to the reference genome. Particularly with reads that originate from a locus that has translocated to the opposite strand. In this case, the strand-specificity is inverted, and the aligner may attempt to map the read to the wrong strand. These false-stranded reads can be detected by the high proportion of GA mismatches on the Watson (+) strand or CT mismatches on the Crick (-) strand. From experience, a threshold of 3-5% regarding the length of the read is usually enough to identify false-stranded reads. As there is a high probability that these reads originate elsewhere in the genome, they are usually excluded from the alignment because it is not a trivial task to infer where the locus may have translocated to (Onishi-Seebacher and Korbel 2011).

Finally, filtering based on multi-mapped reads or PCR duplicates may also be considered, as in standard NGS experiments. Any given sequencing read originates from a fragment derived from a specific position on the genome. During alignment, however, equally-scoring alignments may be possible, particularly in highly repetitive regions or regions of low complexity (Treangen and Salzberg 2011). If each of these alignments is considered separately during quantification of methylation levels, then the chance of error is increased as one of these regions may carry a different methylation state relative to the read. The options are to exclude these alignments entirely, as is performed intrinsically by some sequence aligners (Krueger and Andrews 2011; Guo et al.

2013), or to accept the increased chance of error by selecting one such alignment at random, or including all multiple hits from that read. Regarding PCR duplicates, several tools exist to identify such reads that arise from a single DNA fragment, such as Picard MarkDuplicates (<http://broadinstitute.github.io/picard>) and samtools rmdup (Li et al. 2009). These tools identify PCR duplicates based on the proportionally higher likelihood that identical reads arise from PCR, then that they are separate fragments. In this case, such reads are counted only once during quantification of methylation levels. This deduplication is appropriate only in WGBS sequencing protocols, whereas a high proportion of identical reads are expected in RRBS and target capture. If PCR is absent (or negligible) during library preparation this step should be avoided altogether.

4.5.3 Methylation Calling

To infer the level of DNA methylation at any given cytosine position within the test sample, the methylated bases (cytosine) in the reads overlapping that position are evaluated to give the proportion relative to the total coverage. Non-CT nucleotides are typically ignored, and in practice there is only one major divergence in the implementation of this method by different software: what to do with CT polymorphisms that arise before bisulfite treatment has taken place. Most tools simply ignore evidence of SNPs, counting only reads with either C or T. While it may be appropriate not to count such bases towards the level of cytosine methylation, there may still be methylation present on such nucleotides particularly in the case of adenine (Ratel et al. 2006).

A more robust analysis will recognise that not all read-based thymines overlapping reference-based cytosine positions are representative of the bisulfite treatment (Yaping Liu et al. 2012). A mutation in this context is deceiving, as we might interpret it as demethylation resulting from an epigenetic mechanism rather than a genetic one. The only way to identify such a mutation from the bisulfite sequencing data itself, rather than from independent genotyping of the test sample following conventional sequencing, is to compare the overlapping reads in that position to those on the opposite strand. The DNA strands after treatment with sodium bisulfite are no longer complementary to each other since the methylation information is strand-specific. Artificially converted bases should therefore always lack a complementary base on the opposite strand, whereas a CT mutation will have support on the opposing strand. This difference can be used to exclude such base positions from the analysis; the independent testing of both strands in this manner however would suggest that twice the coverage be required to achieve a similar degree of confidence to standard genotyping. Software such as MethylDackel (<https://github.com/dpryan79/MethylDackel>), for example, attempt to infer the likelihood of a

SNP during methylation calling, by setting a filtering threshold on opposite-strand alignments and skipping the position entirely if there is a user-determined level of support for a SNP.

Further consideration during methylation calling should also be made when using paired end sequencing, to regions that overlap between read pairs. In Illumina sequencing, each read pair will start from opposing ends of a single DNA fragment, and it is possible in cases where the fragment size is less than double the number of sequencing cycles that a part of the fragment will be sequenced twice for the same read pair (Magoč and Salzberg 2011). In this case, methylation information in this overlap region is redundant and does not constitute an independent observation of the methylation status as if they arose from separate DNA fragments. It is therefore wise to identify overlapping regions prior to methylation calling and mask those bases from either one of the two reads of a pair.

4.6 Alternative Methods

Although the use of bisulfite sequencing to analyse DNA methylation in CG, CHG, and CHH context is highly relevant to the study of epigenetics in plant ecology, it is not all-encompassing. There are indeed other techniques based on similar principles that can be used to capture methylation information, such as affinity-based methods: methylated DNA immunoprecipitation (MeDIP), which can be used in combination with high-resolution DNA microarrays or NGS, and methyl-CpG-binding domain (MBD) sequencing.

More recently there has also been some investigation into treatments that can facilitate nucleotide-level base conversion without the harsh side effects of sodium bisulfite (Yibin Liu et al. 2019). In addition, the advent of long-read sequencing technologies such as single molecule real-time analysis with PacBio and ONT has also provided an alternative to bisulfite sequencing. The base calling (described in Chapter 1 section 1.3.1) generates profiles for each nucleotide, which differ between bases with and without base modifications, and can thus be used to differentiate them (Flusberg et al. 2010; Xie et al. 2021). This then has the advantage of detecting DNA methylation without the need for harsh bisulfite treatment, while also allowing for detection of other forms of base modification, using longer reads. Unfortunately, the profiles can prove difficult to make accurate inferences from, but development of machine learning techniques may be a promising avenue of advancement in this regard in the near-future.

5 From Read Alignment to DNA Methylation Analysis

5.1 Introduction

Over the three decades following the conception of bisulfite sequencing by (Frommer et al. 1992) it has become the foundation of many investigations linking cytosine methylation with epigenetics at nucleotide-level resolution. Cytosine methylation is among the most abundant base modification in eukaryotes, involving the addition of a methyl group (CH₃) to the 5th carbon position of the cytosine ring to form 5-methylcytosine (5mC). In model plants and crops, 5mC has been associated with changes in gene expression (Jaenisch and Bird 2003; Xiaoyu Zhang et al. 2006; Lang et al. 2017), chromosome interactions (S. Feng et al. 2014; Grob, Schmid, and Grossniklaus 2014) and genome stability through the repression of transposable elements (Mirouze et al. 2009; Tsukahara et al. 2009). The role of 5mC in epigenetics is well studied in model organisms (see Chapter 1), but with falling sequencing costs and advances in modern sequencing technology there is incentive now to extend this research to non-model species.

As described in previous chapters, the alignment of bisulfite-treated reads to a reference genome is evidently an important step during downstream processing. Standard read mapping tools are not suitable for this type of data however due to the high number of converted bases which would result in alignment errors (see Chapter 4 section 4.5.2). Reduction of reporting error thresholds lead to a high proportion of false positive alignments, so specific tools have instead been developed to explicitly enable read mapping of bisulfite data. Though such tools number at considerably less than the 60 conventional short-read alignment programs identified by (Fonseca et al. 2012) (see Chapter 1 section 1.3.3), a similar assessment by (Tran et al. 2014) identified at least 16 distinct approaches for bisulfite data. Given the wide variety of approaches that therefore exist for mapping bisulfite data, choosing the right tool can be daunting for scientists without formal training in bioinformatics, and is influenced considerably by the context and scope of each study. Previous independent comparisons among such tools have focused on algorithmic differences (Tran et al. 2014), combinations of pre- and post-processing techniques (Tsuji and Weng 2016) or a small

range of tools on model data (e.g. human) (Chatterjee et al. 2012; Kunde-Ramamoorthy et al. 2014). Such reviews help to refine existing computational approaches and aid in new software development, but it is important also to consider the biological implications of emerging applications, such as non-model plant data, in order to establish best-practices for analysis.

Plant genomes are notoriously difficult to work with due to large (De La Torre et al. 2014) and repetitive genome sequences (Mehrotra and Goyal 2014), regions of low complexity, and a variably high degree of zygosity and polyploidy (Wendel 2015). These factors can confound both genome assembly and alignment, often resulting in low-quality genomes with poor contiguity and a large number of misassemblies. With non-model species there is a greater likelihood that the genome will exist in a draft state, thus compounding these problems further in regard to the level of information that can be reliably inferred in downstream analyses. These issues are usually mitigated for example with long-read sequencing technologies, such as PacBio or ONT, which are rarely leveraged to full effect in draft assemblies. In terms of bisulfite sequencing, fragmentation caused by the harsh chemical treatment and issues in resolving accurate base calls on methylated cytosines reduces the viability of long read technology in the present application. Illumina short reads remain the most widely-adopted form of NGS in the study of DNA methylation.

This chapter addresses the use case of non-model plant data for bisulfite read alignment. To this end, a selection of nine bisulfite short-read alignment tools are compared using a combination of real and simulated sequencing data, for three non-model plant species which vary in terms of genome composition and assembly quality (Table 4). These species are represented in the broader initiative of the EpiDiverse consortium (<https://epidiverse.eu/>), and include a high-quality (almost chromosome-level) assembly of the perennial Rosaceae *Fragaria vesca* (Edger et al. 2018) and two fragmented scaffold-level assemblies; one with higher repeat content in the case of the annual Brassicaceae *Thlaspi arvense* (Dorn et al. 2015), and one with lower, in the case of the unpublished, *de novo* assembly of the deciduous tree species *Populus nigra* (currently in development under the EpiDiverse initiative). Each species serves as a representative use case for other non-model organisms. A shortlist of representative software were chosen in-part based on availability through Bioconda (Grüning et al. 2018) (for reproducibility), and according to the extent of software maintenance and adoption under current practices. These include Bismark (Krueger and Andrews 2011), BS-Seeker2 (Guo et al. 2013), BSMAP (Xi and Li 2009), bwa-meth (Pedersen et al. 2014), ERNE-BS5 (Prezza et al. 2012), GEM3 (Marco-Sola et al. 2012), GSNAP (Wu and Nacu 2010), Last (Frith, Mori, and Asai 2012) and segemehl (Otto, Stadler, and Hoffmann 2012, 2014).

Table 4. Basic assembly statistics for non-model plant species referenced in this study

Species	Genome size (Mb)	Scaffolds	Scaffold N50 (Mb)	Repeat content (%)	Accession	Source
<i>F. vesca</i>	220	29	33.9	33	Fragaria-vesca_v4.0.a1	rosaceae.org
<i>T. arvense</i>	343	6,768	0.14	55	GCA_000956625.1	NCBI
<i>P. nigra</i>	417	9,533	9.49	32	unpublished	unpublished

Note: Repeat content is given as a percentage of the total genome space

Read mapping for each tool is evaluated in terms of precision-recall of the simulated bisulfite-treated reads when compared to unique alignments of a corresponding, unconverted data set in each case mapped using the fully sensitive aligner RazerS 3 (Weese, Holtgrewe, and Reinert 2012). Furthermore, methylation profiles are derived from real data and the tools evaluated based on the mean absolute deviation of methylation values, using a subset of difficult-to-map regions where a $\log_2(x) > 1$ absolute deviation in sequencing depth is observed overlapping a repeat annotation in at least one tool. Processing time and peak memory consumption are also measured over incremental levels of sequencing depth to assess the comparative performance of each tool on a standard, representative computing architecture.

5.2 Materials and methods

5.2.1 Reference species

All species are non-model plant organisms selected under the broader initiative of the EpiDiverse consortium. Each reference varies in its overall assembly contiguity and underlying feature complexity (Table 4), representing different stages of assembly completeness. Complex repeat and TE annotations were derived using EDTA (Ou et al. 2019).

5.2.2 Natural accessions

To contrast features common to artificial reads and to infer the effect of read mapping on methylation quantification, one natural accession per species (150 bp long paired-end reads, randomly down-sampled to 20x) was mapped in addition to the simulated data. Methylation profiles were derived for each species by aggregating the methylation calls obtained following read alignment with each tested software. These profiles represent the underlying truth sets for then

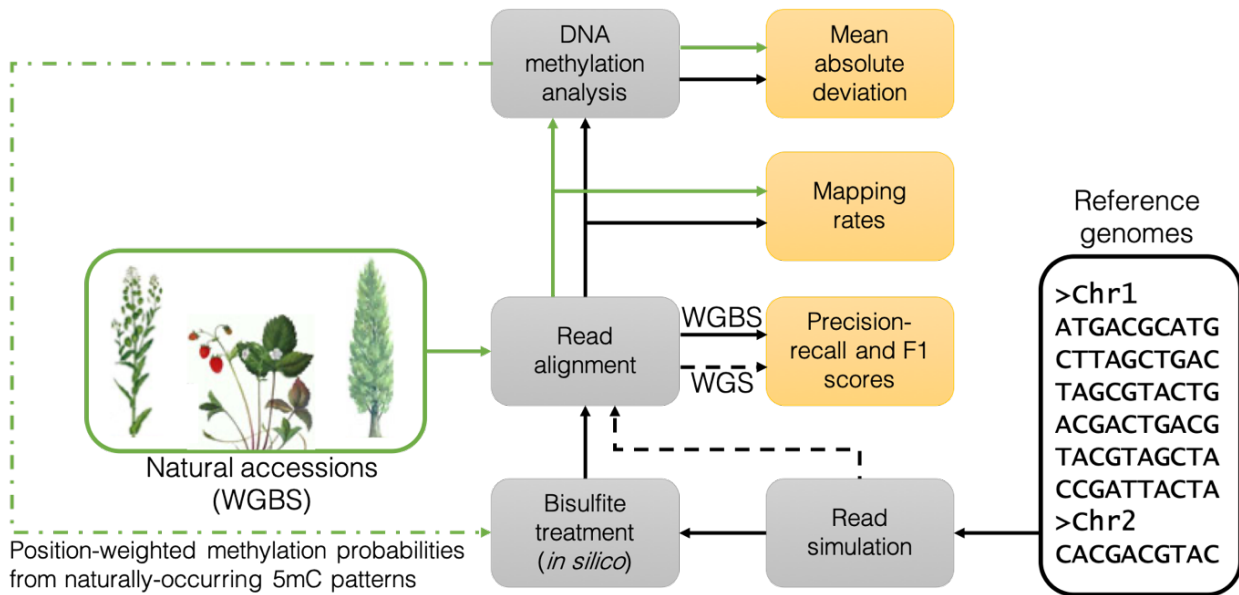


Figure 15. Schematic overview of the experimental design used in this study. Grey boxes illustrate procedures, whereas yellow boxes correspond to benchmarking measures. Processing of simulated and real data is indicated by black and green arrows, respectively. Naturally-occurring 5mC patterns derived from natural accessions are used to inform bisulfite treatment of simulated reads *in silico*. The simulated WGBS data is mapped with each tested aligner and compared to RazerS 3 alignments of untreated reads, in order to evaluate precision-recall. These alignments are also compared to alignments from natural accessions by calculating mapping rates. Methylation profiles from simulated data are compared to the naturally-occurring 5mC patterns they are derived from, in order to evaluate the influence of each software on downstream DNA methylation analysis.

simulating artificial reads based on naturally occurring methylation patterns. Figure 15 describes the overview of interaction between the different datasets.

5.2.3 Read simulation

Five independent sets of 125 bp paired-end reads were generated artificially from each reference genome using the read simulator Sherman v1.7 (<https://www.bioinformatics.babraham.ac.uk/projects/sherman/>). The datasets range incrementally from 1 to 20x sequencing coverage and were generated initially with a variable insert size ranging from 0 to 500, a random nucleotide error rate of 0.5% and a bisulfite conversion rate of 0. A variable length adaptor sequence was also generated, which was subsequently trimmed using cutadapt v2.5 (Martin 2011). The unconverted reads were then processed by an in-house script which applied a random 99% bisulfite conversion rate, yielding in the end two corresponding sets of simulated reads in FASTQ format, with and without bisulfite conversion. An additional set of artificial reads were converted from the 20x dataset in each species, using position-weighted

conversion probabilities derived from the aggregate methylome obtained from the natural accessions.

5.2.4 Read alignment

A total of nine current short-read mapping tools were selected to give a representation of current tools with different alignment strategies (refer to Tran et al. (2014) for further detail), with consideration given only to those with availability through Bioconda (Grüning et al. 2018), in the interest of reproducibility (Table 5). Each software was installed on a server architecture housing 64 cpus with a total of 256 Gb memory (Suppl. Table C.1). For testing purposes, the tools were run with default parameters, which can be interpreted as the best approximation of a “general use case”. Relative processing time (real) and peak memory allocation (resident set size) are reported for each tool, utilising a maximum of eight parallel threads so that results can be relevant to those working e.g. on a laptop or similar. Paired-end data from natural accessions were mapped both in paired-end and single-end mode, after obtaining the reverse complement of mate 2 *in silico*, for comparison of mapping rates.

Table 5. Short-read alignment software tested in this study for mapping bisulfite sequencing reads. Primary alignments are randomly selected from equal-scoring alignments of multi-mapping reads where indicated, and otherwise not reported at all under default parameters.

Software	Version	Default Reporting	Alignment Strategy	Index Structure
Bismark	0.22.3	unique best	3 letter	BWT (bowtie2)
BS-Seeker2	2.1.7	unique best	3 letter	BWT (bowtie2)
Last	1021	unique best	wild card	Spaced suffix array
BSMAP	2.90	unique best / random	wild card	Hash table (SOAP)
BWA-meth	0.2.2	unique best / random	3 letter	BWT (BWA)
ERNE-BS5	2.1.1	unique best / random	wild card	Hash table
GEM3	3.6.1	All-first-N / random	3 letter	Custom FM-index
GSNAP	2019-09-12	All-first-N / random	wild card	Hash table
segemehl	0.3.4	All / random	wild card	Enhanced suffix array

Note: BS-Seeker3 is available but was unable to run successfully on the provided computing infrastructure, and has no recipe in Bioconda at the time of publication.

5.2.5 Mapping rates

Read alignments from each tool were compared in both simulated data and natural accessions (real) data for each species in terms of the overall mapping rate for primary alignments with a minimum mapping quality (MAPQ) threshold of 1. On real sequencing data from natural accessions, mapping rates were calculated additionally for alignments of paired-end data in single-end mode, and also stratified by alignment edit distance (i.e. number of non-bisulfite mismatches) for paired-end alignments. Custom in-house scripting was used to obtain the appropriate edit distance where it was not reported by default by the alignment software.

5.2.6 Precision-recall

Read alignments from each tool were compared i) to the point of origin of the read according to the metadata obtained from the read simulation tool, and ii) to an additional truth set generated by aligning the unconverted reads to the reference with the fully sensitive aligner RazerS 3 (Weese, Holtgrewe, and Reinert 2012), discarding reads that aligned to multiple loci. The higher base complexity in unconverted reads gives an advantage to aligners compared with bisulfite-converted reads. The comparison between the truth set and the bisulfite read alignments allow for the identification of true positives, which demonstrate indirectly the false positives and false negatives (Table 6) derived by each method through the calculation of recall, described in equation (5), and precision, described in equation (6). True positive alignments must occur in the same orientation and with the start coordinate within 5 bp of the corresponding alignment in the truth set. To limit the effect of sampling, the arithmetic means of precision and recall were calculated over all independent simulated datasets (1-20x) for each tool. Tools were then assigned an F1 score, described in equation (7), which reflects the balance of precision and recall through calculation of the harmonic mean of both measures.

Table 6. The confusion matrix indicates the relationship between true positives and false positives. Precision is calculated by taking the true positives as a proportion of the predicted condition positives, and recall is calculated by taking the true positives as a proportion of the true condition positives. The F1 score denotes the harmonic mean of precision and recall.

		True Condition	
		Positives	Negatives
Predicted Condition	Positives	True Positives	False Positives (Type II error)
	Negatives	False Negatives (Type I error)	True Negatives

$$recall = \frac{True\ positives}{True\ positives + False\ negatives} \quad \text{equation (5)}$$

$$precision = \frac{True\ positives}{True\ positives + False\ positives} \quad \text{equation (6)}$$

$$F1\ score = 2 \cdot \left(\frac{recall \cdot precision}{recall + precision} \right) \quad \text{equation (7)}$$

5.2.7 Coverage deviation

Regions of \log_2 -fold differential sequencing depth were calculated for each tool in comparison to unique RazerS 3 alignments using deepTools v3.4.3 bamCompare (Ramírez et al. 2016), after filtering bisulfite alignments based on a minimum MAPQ threshold of 1. The representation of such regions in the genome space of repeat annotations is analysed with a Fisher test implemented by bedtools v2.27.1 fisher (Quinlan and Hall 2010). Regions with a minimum absolute deviation in sequencing depth of $\log_2(x) > 1$ in at least one tool are intersected with repeat annotations using bedtools v2.27.1 intersect (Quinlan and Hall 2010), to identify a “difficult-to-map” subset of the genome space for comparative DNA methylation analysis.

5.2.8 DNA methylation analysis

Methylation profiles for both natural accession data and artificial data were derived in all methylation contexts (i.e. CG, CHG, CHH) using MethylDackel v0.5.0 (<https://github.com/dpryan79/MethylDackel>). The tool adjusts for overlapping paired-end reads, and can account for methylation bias at the 5-end arising during library preparation due to unconverted nucleotides incorporated by end-repair. All alignments were filtered based on a minimum MAPQ score of 1, and positions with a minimum base quality of 1. The methylation calls from natural accession data, produced following alignment with each of the tested software, were combined into an aggregate methylome for use during read simulation of artificial data to confer position-weighted conversion probabilities from naturally occurring 5mC patterns. Resulting methylation calls from the simulated data, produced after aligning with each of the tested software, were then compared back to the aggregate methylation profile over the difficult-to-map regions to evaluate the methylation differences in terms of mean absolute deviation.

5.3 Results

Precision-recall profiles derived from simulated read alignments demonstrate higher F1 scores when comparing to equivalent, unconverted alignments obtained from RazerS 3 (Figure 16), but follow a similar behaviour in terms of dataset difficulty when comparing to the biological point of origin (Figure 17), suggesting that the underlying feature complexity of each genome tested does not deter mapping beyond what can be expected from standard Illumina paired-end sequencing data. When filtering alignments by a minimum MAPQ threshold of 1, the aligners BSMAP and BWA-meth consistently exhibit the highest F1 scores across all datasets, followed closely by Bismark, GEM3 and Last.

Despite a relatively high repeat content relative to the genome space and a highly fragmented assembly, *T. arvensis* perhaps represents the most straightforward simulated dataset in this benchmark, since artificial reads originate only from within scaffolds so they have fewer potential loci to map back to. Conversely, *F. vesca* appears to be the most difficult despite its completeness and relative size. Comparisons with real data demonstrate lower mapping rates overall (Figure 18), particularly in less contiguous and less polished assemblies, possibly due in-part to the presence of discordant reads overlapping break points between scaffolds. Bismark and BS-Seeker2 appear to be particularly susceptible to this, which can be unveiled by aligning the data in single-end mode (Figure 19). The remaining gap can be largely explained by the fact that neither tool seems to output read alignments with more than four to five errors relative to other tools (Figure 20). Taken together it results in fewer methylation calls for both of them (Figure 21), which could potentially confound downstream methylation analysis.

As the difficulty of each dataset increases each tool tends to maintain a level of precision at the expense of recall, whereas GSNAP seems to traverse along the vector of $y = x$, and segemehl appears to struggle initially with the *T. arvensis* dataset perhaps in-part due to the highly fragmented nature of the reference. The aligners GEM3 and BSMAP tended to be among the most sensitive, except for the *F. vesca* dataset where GSNAP also recovered a greater proportion of positive alignments. The lowest recall was observed consistently for ERNE-BS5, which appears to apply a non-standard usage of MAPQ by “binning” alignments either at MAPQ=0 or MAPQ=60. This is reflected by a comparatively high precision relative to the other tools, similar to Bismark and bwa-meth. Further refinement of alignments in other tools by filtering MAPQ thresholds would likely result in improved levels of precision at the cost of recall, with the exception of BSMAP which

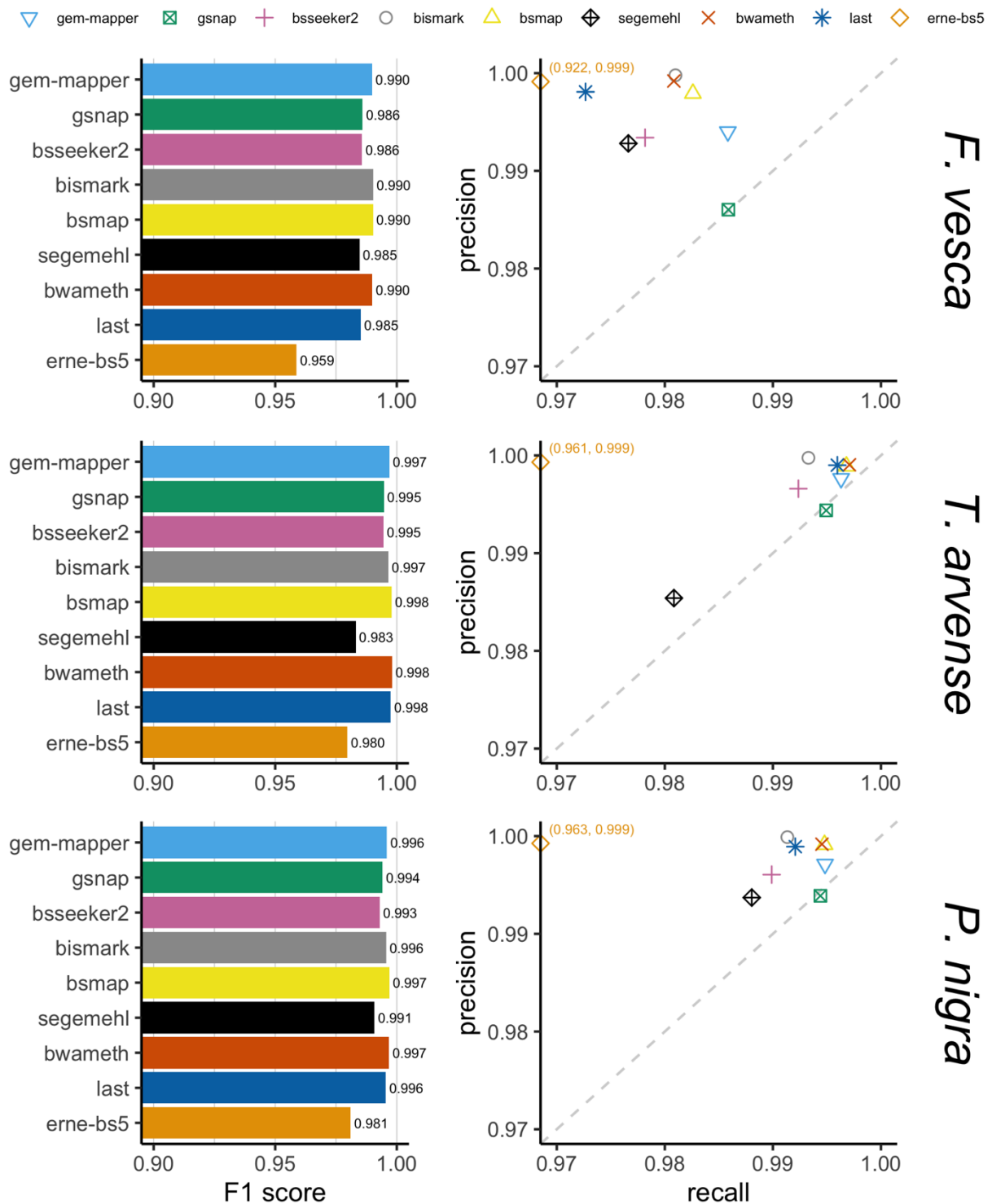


Figure 16. F1 scores and precision-recall for simulated reads mapped by each aligner, as determined by the alignment of equivalent, unconverted reads by RazerS 3, demonstrating the response trade-off at close to maximum recall with a minimum mapping quality (MAPQ) threshold of 1. BS-Seeker2 and BSMAP do not make use of MAPQ scores, and ERNE-BS5 partitions alignments either at MAPQ = 0 or MAPQ = 60. The F1 score is the harmonic mean of precision and recall, which reflects the ranking of each tool relative to the overall balance of both measures. In the right-hand panels, ERNE-BS5 in each case falls out-of-bounds and is annotated with the appropriate coordinate (recall, precision).

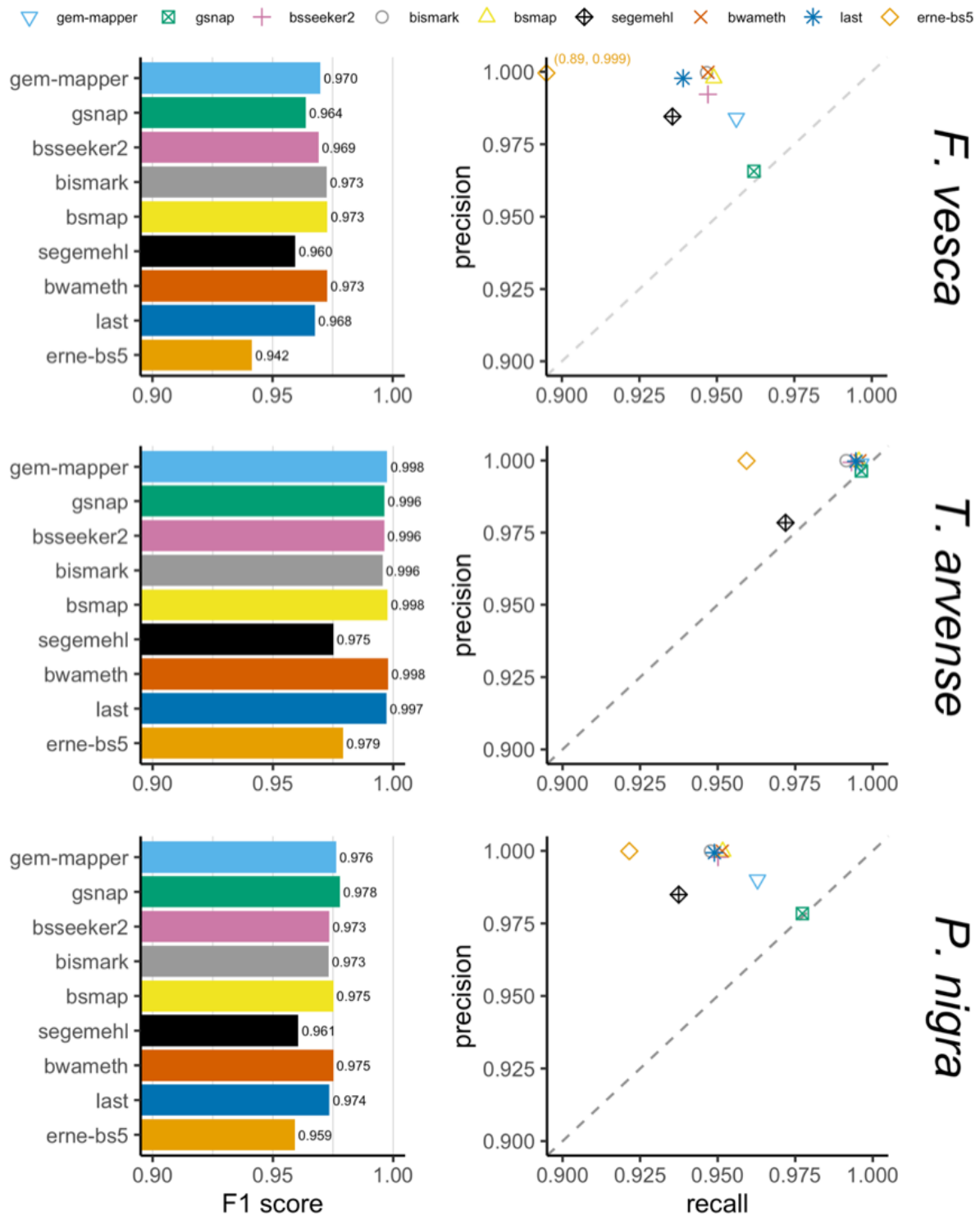


Figure 17. F1 scores and precision-recall for simulated reads mapped by each aligner, as determined by the known biological point-of-origin of reads according to the read simulator, demonstrating the response tradeoff at close to maximum recall with a minimum mapping quality (MAPQ) threshold of 1. BS-Seeker2 and BSMAP do not make use of MAPQ scores, and ERNE-BS5 partitions alignments either at MAPQ = 0 or MAPQ = 60. The F1 score is the harmonic mean of precision and recall, which reflects the ranking of each tool relative to the overall balance of both measures. In the right-hand panel for *F. vesca*, ERNE-BS5 falls out-of-bounds and is annotated with the appropriate coordinate (recall, precision).

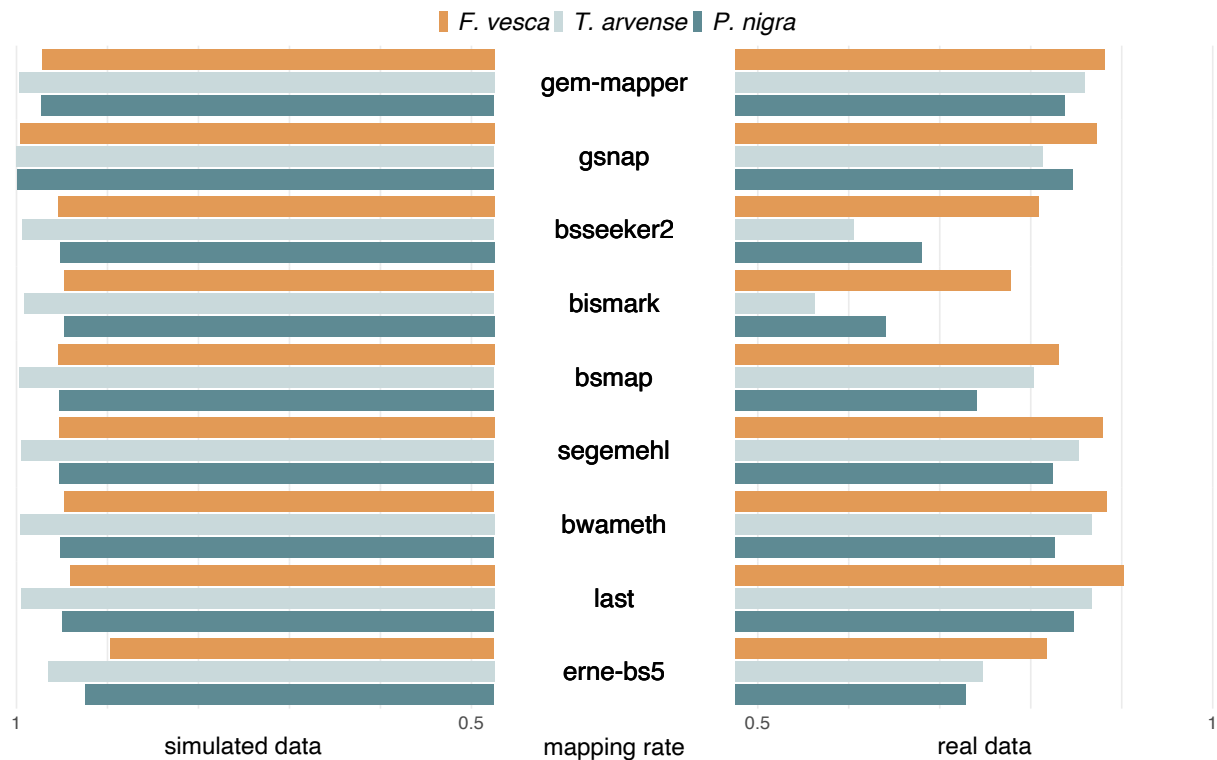


Figure 18. Mapping rates of short-read aligners. Comparisons between simulated and natural accession (real) data for each test species and each tool, given a minimum mapping quality (MAPQ) of 1. Reads from simulated data are generated from each corresponding reference genome and thus expected to behave concordantly, with little sequence variation and minimal influence of base quality, whereas real data may be subject to discordant alignments arising for example from poor reference contiguity and/or genomic rearrangement.

does not make use of MAPQ. Given a minimum MAPQ threshold of 1, the aligners segemehl and GSNAP scored lowest in terms of overall precision.

Regions with an absolute deviation of sequencing depth of $\log_2(x) > 1$ in at least one tool represent a total of approximately 9.7 Mbp, 1.2 Mbp and 16.4 Mbp of the total genome space (4.39%, 0.34% and 3.92%), respectively, in *F. vesca*, *T. arvensis* and *P. nigra*, whereas complex repeat annotations (computed with EDTA) comprise approximately 73.4 Mbp, 190.1 Mbp and 135.2 Mbp. Independent F-tests of the resulting intersection overlaps, in each species, indicate they are overrepresented in the genome space ($P < 1.0 \times 10^{-6}$) at approximately 8.3 Mbp, 1.0 Mbp and 16.4 Mbp (3.75%, 0.30% and 2.11%). These regions are considered “difficult-to-map”, and the difference between the alignment tools relative to RazerS 3 is reflective of how multi-mapping reads are handled in relation to MAPQ (Figure 22).

In all cases, it is expected that mean absolute deviation is inversely correlated with sequencing depth, as a greater number of overlapping reads should reduce the impact of spurious alignments.

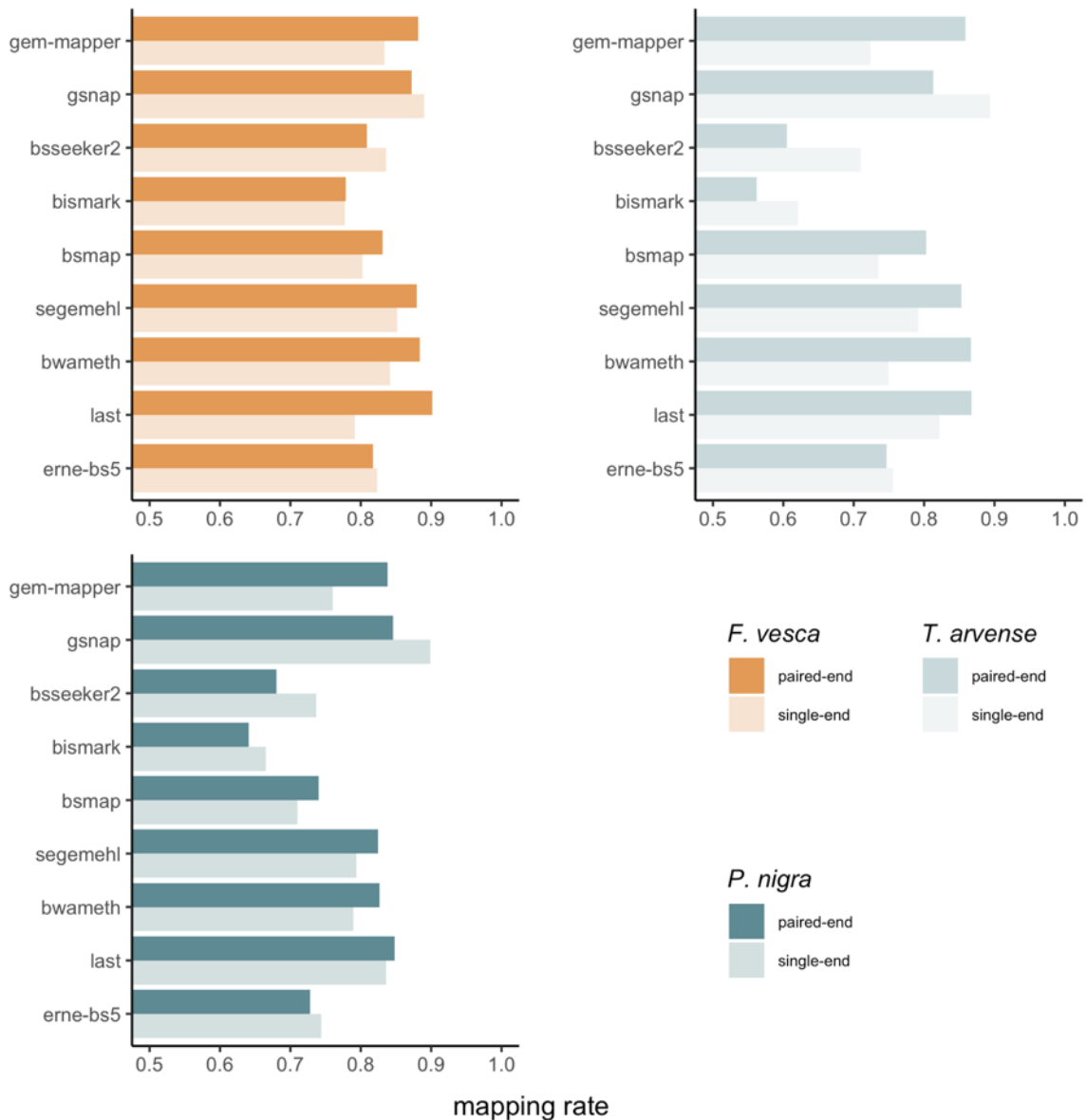


Figure 19. Comparison of mapping rates with each software after aligning the same paired-end WGBS reads from natural accession (real) data in each species, in both paired-end mode and single-end mode. Most tools achieve marginally higher mapping rates in paired-end mode, with the exception of Bismark, BS-Seeker2, ERNE-BS5 and GSNAP. In particular, both Bismark and BS-Seeker2 appear to lose sensitivity in paired-end mode in correlation with an increasing level of fragmentation in the corresponding genome assemblies.

For some tools however the absolute deviation increases again for higher values of minimum sequencing depth in difficult-to-map regions, particularly in the range of $>10x$ where the per-strand depth is greater than the expected mean (Figure 22). This indicates a tendency to map reads which likely differ in their point of origin, thus apparent to some extent in all software with ‘All’ or ‘All-First-N’ reporting strategies for multi-mapping reads, and additionally ERNE-BS5 (random best) and Last (unique only). The influence of such alignments from these tools may be curtailed by setting upper limits for sequencing depth or by more stringent filtering on MAPQ.

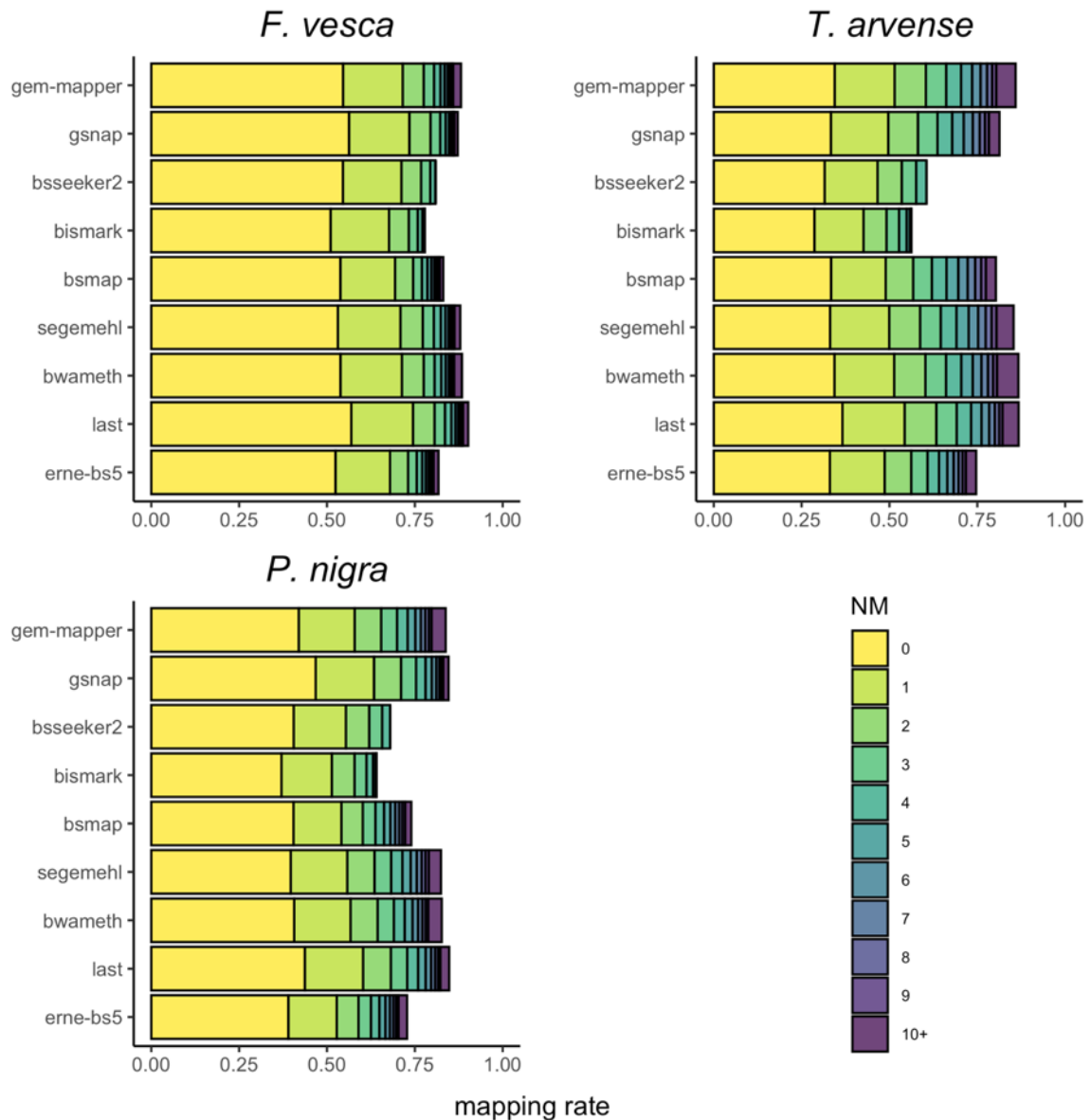


Figure 20. Comparison of mapping rates with each software for paired-end alignments from natural accession (real) data in each species, stratified by the number of alignment errors (NM) which cannot be attributed to the treatment with bisulfite. Most of the aligners allow for up to ten or more errors in the read alignments below a certain number of errors. The exceptions are Bismark and BS-Seeker2 which appear to have a soft-/hard-threshold at 4-5 errors per read.

Comparisons of the mean deviation in methylation rate over all positions, as a function of a threshold on the minimum sequencing depth within difficult-to-map regions, indicate that all software (with the exception of ERNE-BS5) differ only marginally from the expected methylation rate in natural accessions (Figure 23), at lower depth thresholds, regardless of the recovered fraction of independent sites that are called (Figure 22). A higher rate indicates a potential preference towards aligning methylated reads, which could have implications for downstream methylation

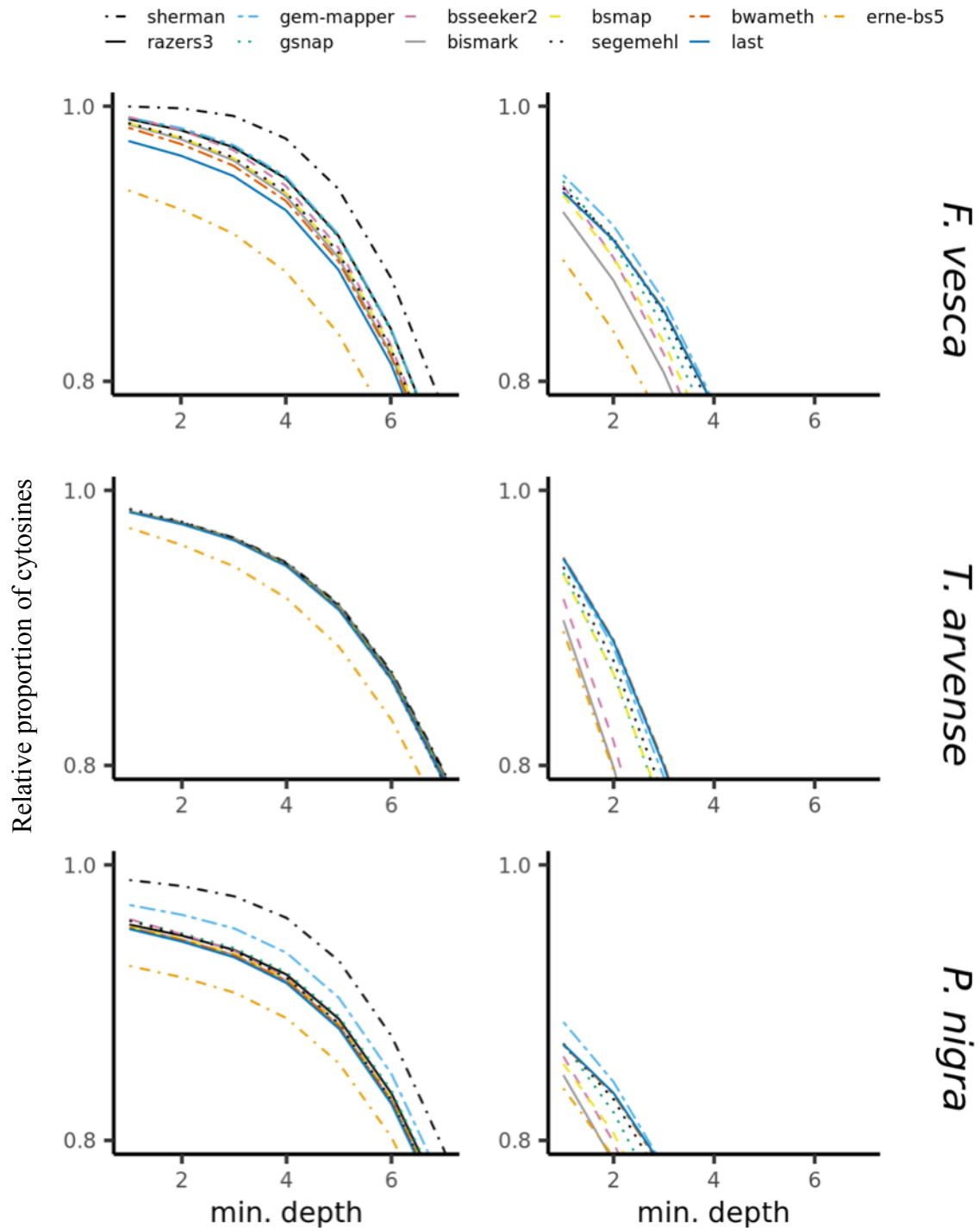


Figure 21. Global methylation site dropout. Total proportion of genomic cytosines in all methylation contexts (i.e. CG, CHG, CHH) derived from each aligner in response to varying the minimum sequencing depth threshold. The expected mean strand-specific sequencing depth is 10x. The left-hand panels represent simulated data, whereas the right-hand panels represent natural accession (real) data.

analysis in such regions. The tendency is not apparent when considering the global methylation profile across the whole genome.

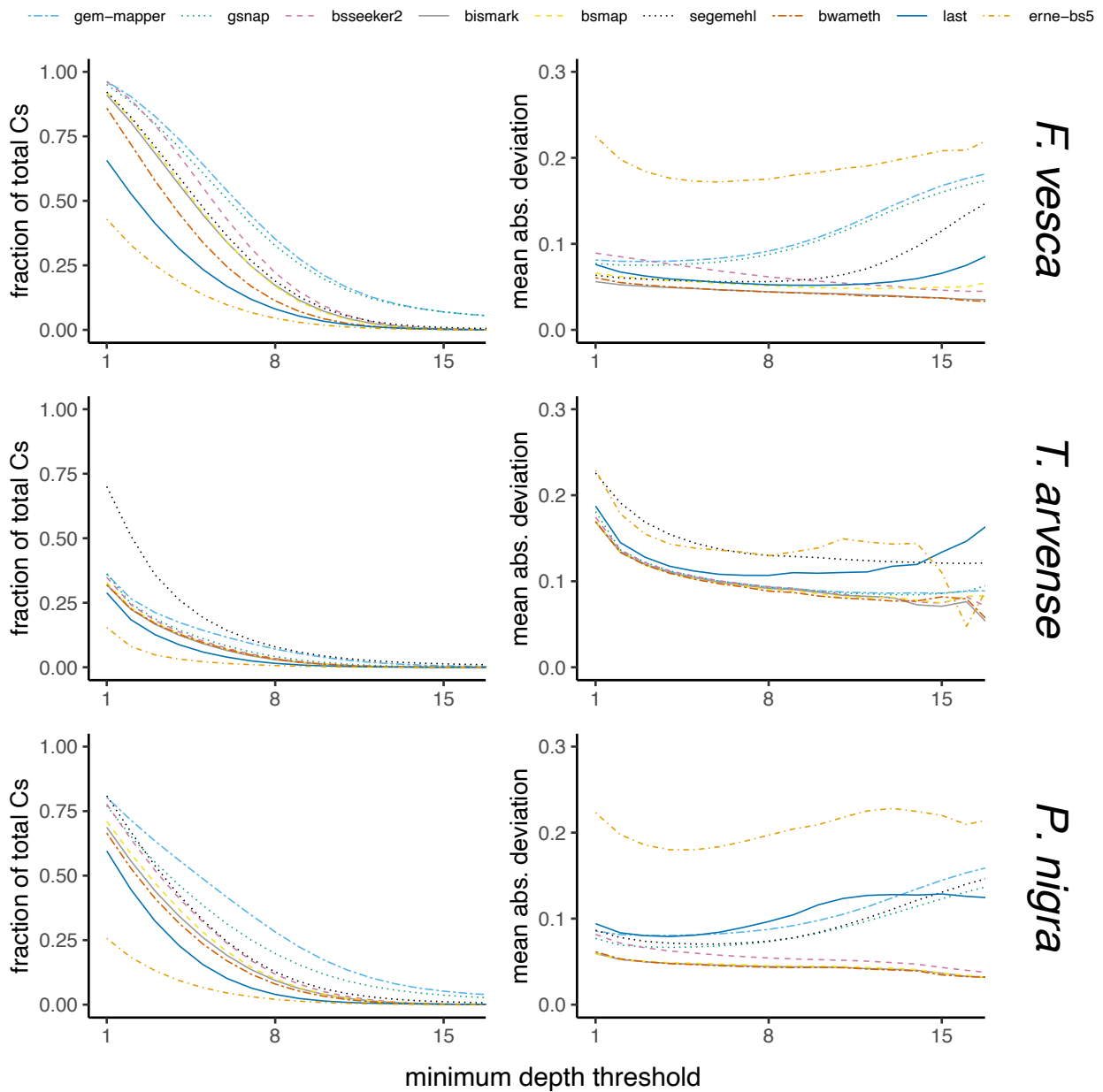


Figure 22. Fraction of total cytosines and mean absolute deviation of methylation calls. Comparisons between tested software in terms of the methylation profiles derived from simulated data, in all methylation contexts, over difficult-to-map regions encompassing $\sim 3.75\%$ of the genome space in *F. vesca*, $\sim 0.3\%$ in *T. arvense* and $\sim 2.11\%$ in *P. nigra*. All plots refer to profiles derived from artificial data simulated based on naturally-occurring 5mC patterns from the corresponding natural accession data. The left-hand panels show the fraction of total cytosines in difficult-to-map regions that are covered by each tool. The right-hand panels show the mean absolute deviation, demonstrating how well the methylation patterns were preserved following alignment, in comparison to the original methylation profiles from natural accession data.

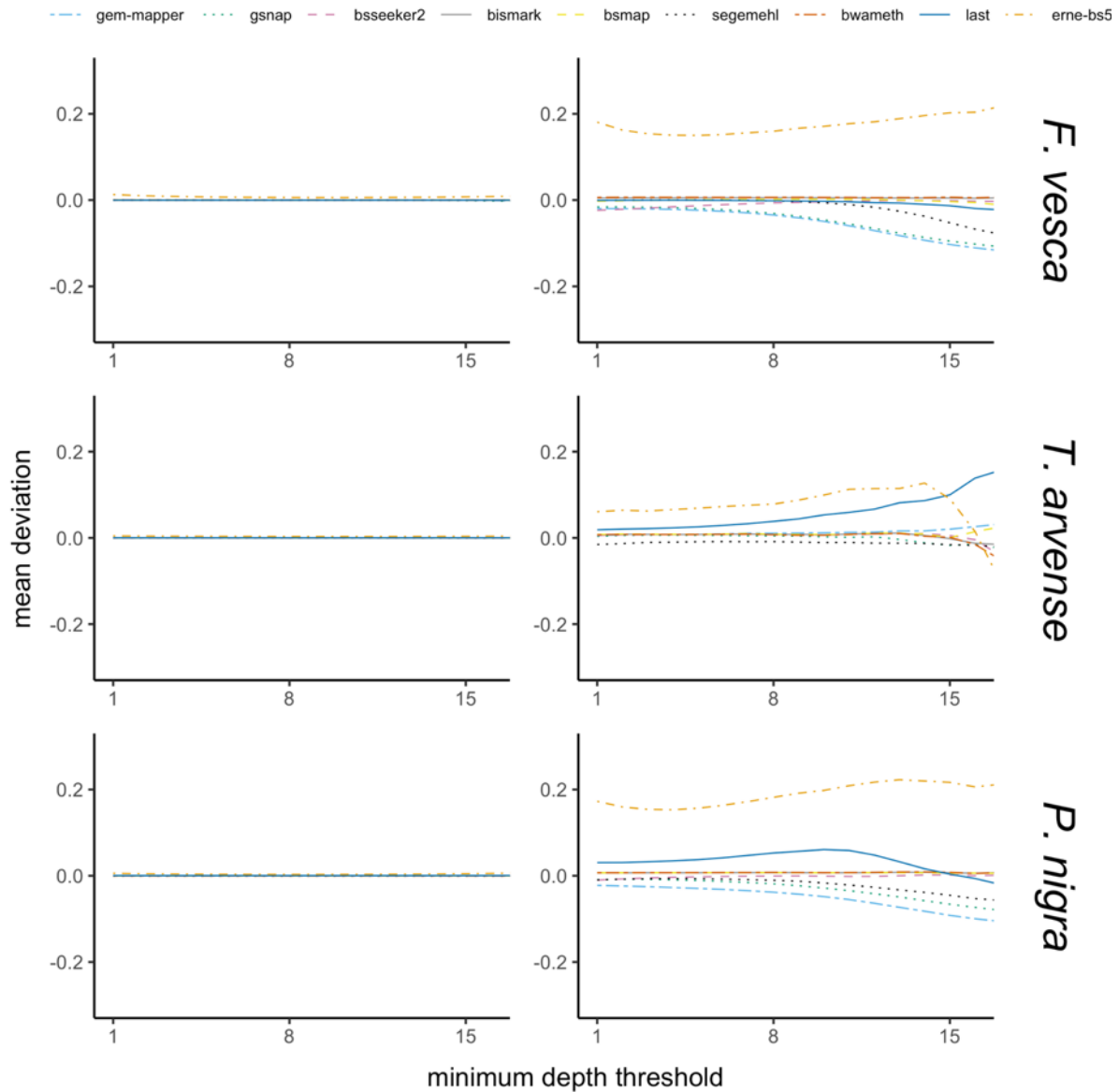


Figure 23. Mean deviation of cytosine methylation (all contexts) in simulated data, relative to the naturally-occurring 5mC patterns derived from natural accessions, as a function of a threshold on the minimum sequencing depth after aligning to each species with each tool. The left-hand panels show the deviation on a genome-wide (global) scale whereas the right-hand panels show only the subset of difficult-to-map regions. Most software perform similarly, but few tend to underestimate the methylation level in difficult-to-map regions at higher-than-expected levels of strand-specific sequencing depth (e.g. > 10x). In contrast, ERNE-BS5 appears to overestimate the methylation level in difficult-to-map regions but otherwise does not differ noticeably from the global methylation level.

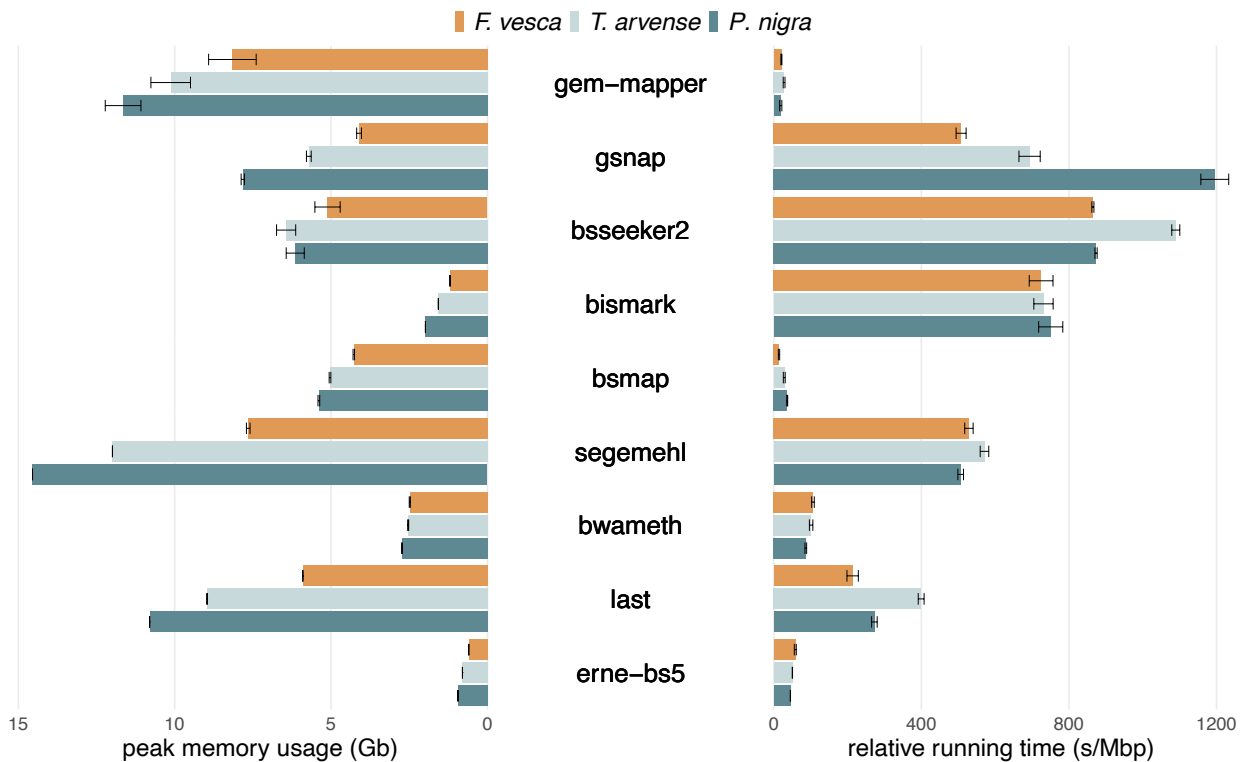


Figure 24. Peak memory and running time on alignments of simulated reads at varying levels of sequencing coverage (1-20x). Peak memory usage is given in terms of resident set size (Gb) and running time in terms of seconds per Mbp for comparison. Memory is dependent on the size of the genome relative to the effect on the index data structure, whereas time is dependent on the total quantity of reads to align. Larger error bars indicate memory usage differences that arise due to differences in sequencing depth, or non-linear increases in process running time.

Finally, when comparing the computation performance, the aligners BSMAP, BWA-meth, ERNE-BS5 and GEM3 exhibited the fastest running times, while BWA-meth and ERNE-BS5 also ran with the lowest demand on peak memory alongside Bismark (Figure 24). For production environments with a focus on high throughput, aligners such as BWA-meth and ERNE-BS5 might be preferred. If computational resources are not a factor, then on-balance BWA-meth and BSMAP are able to make the most of the data available, depending on whether further refinement by MAPQ is required. For non-model data specifically, further consideration might also be given to how discordant alignments are handled by each tool.

5.4 Discussion

Previous studies have shown the imperative to consider methodological differences in the context of downstream methylation analysis, for example when detecting bias in WGBS library preparation strategies (Olova et al. 2018). When mapping bisulfite-converted short reads, prioritising one of either recall or precision might be appropriate when assessing individual alignments, but can lead

to bias in methylation rates. Deriving the correct result over a given position is dependent on maintaining the correct ratio of methylated and unmethylated cytosines from the pool of reads obtained from a given biological sample. This ratio is disturbed not only by inaccurate mapping, as can be more prevalent in software with lower precision, but also by over-filtering alignments based on measures such as MAPQ, as may be prevalent in software with lower recall. The trade-off is more apparent when considering the stringency for handling multi-mapping reads in each tool with respect to MAPQ, particularly over difficult-to-map regions with local minima or maxima in overall sequencing depth.

Adjusting methylation rates or providing confidence intervals based on the evaluated mappability of reference regions might be beneficial for downstream analysis; however, existing tools based on self-alignments of k-mers (Karimzadeh et al. 2018) may overestimate the mappability of heterozygous loci and/or scaffold boundaries in highly fragmented genomes. Furthermore, differences in mean methylation patterns between different software indicate preferences in some instances for mapping methylated loci which are not explained by sequencing depth bias arising through library preparation. More densely methylated reads benefit from increased sequence complexity, which may confer an advantage during read alignment which has a downstream impact on methylation rate. The performance of bisulfite read alignment software is responsive to achieving an optimal balance of precision-recall with respect to both methylation status and the mappability of genomic regions.

It is important to consider that the metrics typically used in benchmarking approaches (e.g. precision, recall, F1 score) tend to reflect only the descriptive statistics of individual cases; they do not account for the full breadth of potential variation between different species, for example. Though model species are often used to make predictions, a more robust statistical approach would strictly be necessary in order to develop a high-confidence model that carries over to other, non-model organisms. In the present context, the benchmarking of software using their default parameters appears most fair as an approximation of a “general use case” and also trivial for any educated user to carry over to other scenarios. Parameter optimisation is dependent on consistent implementation and reproducible behaviour between use cases, and it is a lot to expect from an educated user to select optimal settings for each tool without assistance from an expert. In summary, this study expands upon existing work in assessing current bisulfite alignment software by incorporating a range of emerging applications and shifting focus towards downstream

methylation analysis; however, further refinement is encouraged on a case-by-case basis both in terms of software selection and the optimisation of parameter settings to further improve results.

5.5 A pipeline for WGBS analysis

Based partially on the results of this benchmarking analysis, a computational pipeline for performing alignment of whole genome bisulfite sequencing data was developed for use by the EpiDiverse consortium in the study of epigenetics in plant ecology. It comprises one aspect of the “EpiDiverse Toolkit”, and makes use of common file formats and standards to facilitate interoperability both with other pipelines in the toolkit and external software. The pipeline is implemented with Nextflow under the DSL v2 framework, and facilitates processing of population-scale data in a highly-parallel manner. The system is portable and able to be easily-configured for different computational architectures, with very little bioinformatic expertise required on behalf of the end-user. The software is available at <https://github.com/EpiDiverse/WGBS>

The workflow processes raw data from FASTQ inputs (FastQC, cutadapt), aligns the trimmed reads (ERNE-BS5 or segemehl), and performs extensive quality control and post-processing on the results using custom scripts. The pipeline is able to process sample data based on read groups, and performs estimation of sample-specific non-conversion rates based on either lambda phage “spike-in” or using a given scaffold (e.g. chloroplast). Picard MarkDuplicates is used to remove PCR duplicates computationally, and finally methylation calling and M-bias correction is performed with MethylDackel.

6 There and Back Again: Inferring Genomic Information

6.1 Introduction

Previous chapters have addressed DNA methylation (Chapter 1 section 1.2.2) and its relevance as an epigenetic mark in plant ecology, in addition to the NGS technologies and approaches used to study it (Chapters 1, 4). One important consideration which is often overlooked during the analysis of bisulfite sequencing data, however, is that researchers often need to frame the results of their investigations into epigenetic mechanisms and processes in the context of the genetic background. Common approaches involve for example the use of clones, in order to control for genetic variability and ensure that experimental results correlate only with the tested epigenetic factors. As the scope and scale of such experiments extends to non-model plant species, however, those with different modes of reproduction have to be accounted for and it's not always possible to limit genetic effects experimentally. It is increasingly the case that genetic markers such as SNPs have to be evaluated in parallel, under the same experimental conditions in which studies in epigenetics are performed, necessitating for example the use of conventional sequencing libraries alongside bisulfite sequencing data. This can be prohibitive both in terms of cost and the resulting trade-off in e.g. sequencing depth / number of samples in order to meet statistical power requirements and generate reliable results, especially in consideration of population-level experiments.

Alternatively, though it is not yet feasible with conventional tools, it is nevertheless possible to obtain genetic variants directly from the bisulfite sequencing data itself, owing to the non-complementary nature of opposite-strand alignments in the case of artificial mutations caused by the bisulfite treatment. Conventional tools for variant calling currently have no way to distinguish thymine mismatches arising as a result of natural mutation, which are obscured by the bisulfite conversion and risk being mistaken simply as unmethylated cytosines. More specifically, the variant caller will interpret every artificial mutation throughout the genome as potential evidence of a SNP, skewing the balance of alleles required for accurate genotyping and resulting in many false positives. In order to differentiate true positives from false positives, then, it is necessary to derive allele counts on a per-strand basis prior to variant calling. A natural mutation in a bisulfite context (i.e.

C>T) will retain its true complementarity on the opposite strand, whereas an artificial mutation will be reflected by the complementary base of the original cytosine nucleotide. Such a method can potentially make it feasible to sequence only bisulfite libraries, and yet derive both epigenetic data as well as the genetic variants relevant to a more comprehensive analysis.

Previous attempts to resolve such confounding positions in the genome, to determine both the correct methylation level and reveal underlying SNPs, have resulted in the development of specialised software such as BISCUIT (<https://github.com/huishenlab/biscuit>), Bis-SNP (Yaping Liu et al. 2012), BS-SNPer (Gao et al. 2015), gemBS (Merkel et al. 2019) and MethylExtract (Barturen et al. 2013). Each case combines methylation calling and variant calling into a single, concurrent analysis to produce output in a custom variant call format (VCF). No single approach however considers the variant calling itself as a primary, independent outcome. Users looking additionally to leverage SNP data for e.g. genotyping or purposes unrelated to DNA methylation are therefore limited by the scope and rationale behind the development of existing tools where the priority is to establish methylation. Instead, the present application aims to abstract variant calling as a standalone objective, in order to facilitate analysis with conventional software, such as GATK (McKenna et al. 2010), Freebayes (Garrison and Marth 2012), or Platypus (Rimmer et al. 2014), thereby optimising precision-sensitivity during SNP discovery and allowing users to make the most out of their bisulfite sequencing data for a broader range of purposes.

Under a simple Bayesian framework to variant calling, the conditional probability of observing the true genotype G given the variants observed in the sequencing data D can be represented for example by equation (4) (Chapter 1 section 1.3.4), which formulates the problem as the derivation of a prior estimate of the genotype $P(G)$ and the likelihood of observing the data $P(D|G)$.

Given that NGS data is seldom error-free, even the simplest model will typically incorporate base quality (BQ) information directly into the Bayesian inference of genotypes as a fundamental scaling factor for the data likelihood estimation. The BQ score itself is a phred-based quality value which denotes on each position the estimated probability that the base caller identified the correct nucleotide during sequencing. In the context of variant calling from bisulfite-treated NGS data, any potential nucleotide conversions present in the resulting sequencing reads can, in principle, be considered analogous to zero-quality base calls. Leveraging this mechanism imposes an indirect strand-specificity on potential variants which cannot otherwise be dissociated from the effect of

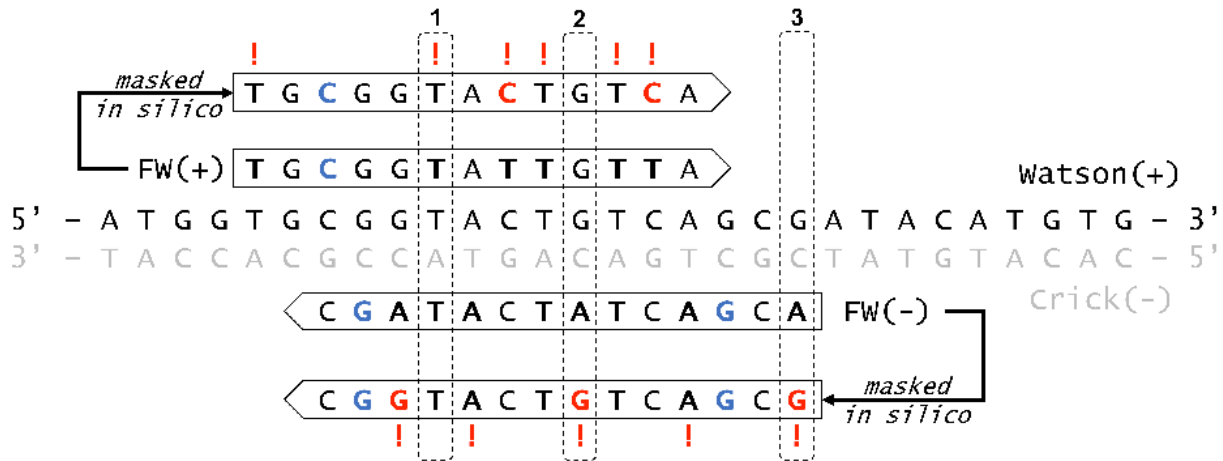


Figure 25. Overview of the double-masking procedure. The central sequence represents the reference genome, with example alignments (+FW and -FW) adjacent to each originating strand. Black, emboldened nucleotides potentially arise from bisulfite treatment. Blue colouring indicates 5mC. Red colouring represents *in silico* nucleotide manipulation, and corresponding base quality manipulations are indicated with an exclamation mark. In example (1) the variant caller is informed only by the -FW alignment, and in (2) only by the +FW alignment. As there is no equivalent Watson (+) alignment in (3) it is impossible to determine whether the apparent G>A polymorphism arises from bisulfite or by natural mutation.

bisulfite conversion, dictating that they be informed only by opposite-strand alignments where the original, complementary nucleotide is hence unaffected by the treatment.

6.1.1 Implementing a new approach

The method presented herein involves a simple “double-masking” procedure which manipulates specific nucleotides and BQ scores on alignments from bisulfite sequencing libraries (Figure 25), with the formal procedure on individual alignments described in Algorithm (1). It involves two steps which are performed *in silico*. First, specific nucleotides in bisulfite contexts are converted to the corresponding reference base, in order to prevent any preselection of sites which are informed exclusively by the artificial bisulfite treatment. This circumvents what can potentially be millions of positions from even being considered by the variant caller as candidate variants for analysis, thus reducing valuable analysis time and conserving computational resources. Second, any given nucleotide which may potentially have arisen due to bisulfite conversion is assigned a base quality (BQ) score of 0. This drives the variant caller to make the correct decision in regards to genotype on positions where there is real evidence of a SNP. As the procedure is informed by decisions made during alignment, it behaves in exactly the same manner and is applicable to both directional and non-directional sequencing libraries. In paired-end sequencing, the procedure applies in a C>T context on mate 1 alignments to the Watson strand (FW+) and mate 2 alignments to the Crick strand (RC-), whereas mate 1 alignments to the Crick strand (FW-) and mate 2 alignments to the

Algorithm 1 The double-masking procedure as performed on each alignment.

```

1: procedure DOUBLEMASKING( $M, W, S$ ) ▷ boolean tests  $M$  for Mate 1 and  $W$  for Watson strand, and a set  $S$  of
   all aligned base pairs
2:    $A \leftarrow \emptyset$ 
3:   if ( $M = true$  and  $W = true$ ) or ( $M = false$  and  $W = false$ ) then
4:      $CT \leftarrow true$  ▷ bisulfite conversion in C>T context
5:   else
6:      $CT \leftarrow false$  ▷ bisulfite conversion in G>A context
7:   end if
8:   for all  $U \in S$  do
9:      $(U_0, U_1, U_2) \leftarrow U$  ▷ each aligned pair  $U$  is a subset containing the corresponding reference base, query
   base, and query base quality, respectively
10:    if  $CT = true$  and  $U_0 = cytosine$  and  $U_1 = thymine$  then
11:       $U_1 \leftarrow cytosine$ 
12:       $U_2 \leftarrow 0$ 
13:    else if  $CT = false$  and  $U_0 = guanine$  and  $U_1 = adenine$  then
14:       $U_1 \leftarrow guanine$ 
15:       $U_2 \leftarrow 0$ 
16:    else if ( $CT = true$  and  $U_1 = thymine$ ) or ( $CT = false$  and  $U_1 = adenine$ ) then
17:       $U_2 \leftarrow 0$ 
18:    end if
19:     $A \leftarrow A \cup \{U\}$  ▷ modified or unmodified pair is added to a new alignment set
20:  end for
21:  return  $A$ 
22: end procedure

```

Watson strand (RC+) follow G>A context. Reads obtained from single-end sequencing behave in an equivalent manner to mate 1 in paired-end sequencing.

In contrast to previous approaches with bisulfite data, the method is applied as a pre-processing step prior to variant calling, thereby facilitating interoperability with conventional, state-of-the-art variant calling software. For validation, SNPs derived from published, experimental whole genome bisulfite sequencing (WGBS) data in human (NA12878) and *Arabidopsis thaliana* (Cvi-0) accessions are compared to high-quality variant standards and high-confidence regions obtained from the NIST Genome in a Bottle initiative (Zook et al. 2014) and the 1001 genomes project (1001 Genomes Consortium et al. 2016), respectively. The method presented herein has been implemented as a standalone python script available at <https://github.com/bio15anu/revelio>, which is intended to be adapted and “plugged-in” to any variant pipeline working with bisulfite data so that the user can choose whichever alignment and variant calling software best suits their purposes. An open-source example of a working pipeline for whole genome data is available at <https://github.com/EpiDiverse/SNP> which is itself a branch

of the EpiDiverse Toolkit (Nunn et al. 2021). The software is also implemented by epiGBS2 in the analysis of reduced-representation bisulfite data (Gawehns et al. 2022).

6.2 Materials and methods

6.2.1 Validation datasets

All datasets analysed in this study are derived from published, public domain resources. High-quality reference variant datasets for *A. thaliana* (Cvi-0) and human (NA12878) accessions were obtained from the 1001 genomes project (<https://1001genomes.org/data/GMI-MPI/releases/v3.1/>) and Genome in a Bottle (GIAB) (https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/NA12878_HG001/NISTv4.2.1/), respectively. The corresponding reference genomes TAIR10 (GCF_000001735.3) and GRCh38 (GCF_000001405.26) were obtained from NCBI. Equivalent WGBS data were obtained from the NCBI Sequence Read Archive under accessions SRX248646 (single-end, ~34X) and SRX3161707 (paired-end, ~46X). Please refer to (1001 Genomes Consortium et al. 2016) and (Suzuki et al. 2018) for further technical specifications regarding these datasets. The original whole genome sequencing (WGS) data for *A. thaliana* Cvi-0 was also obtained, under accession SRX972441 (paired-end, ~62X). Both trimmed reads and alignments from this accession were subject individually to *in silico* bisulfite treatment (~99% conversion rate), using custom in-house python scripts, to generate corresponding, simulated WGBS datasets.

6.2.2 Read processing and alignment

Reads were assessed with FastQC v0.11.8 (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc>) and, where appropriate, trimming performed with cutadapt v2.5 (Martin 2011). WGS alignments were carried out with BWA v0.7.17-r1188 (Li and Durbin 2009), and WGBS alignments with bwa-meth v0.2.2 (Pedersen et al. 2014). Read groups were merged with SAMtools v1.9 (Li et al. 2009), where appropriate, and PCR duplicates subsequently marked with Picard MarkDuplicates v2.21.1 (<http://broadinstitute.github.io/picard>).

6.2.3 Variant calling

Following the double-masking procedure, variants were called using GATK v3.8 UnifiedGenotyper (McKenna et al. 2010), Freebayes v1.3.1-dirty (Garrison and Marth 2012), and

Platypus v0.8.1.2 (Rimmer et al. 2014), in all cases with a hard filter of 1 on both minimum mapping quality (MAPQ) and BQ. Variants were called in addition using Platypus on assembly-mode with $BQ \geq 0$. For comparison, variants from the original bisulfite alignments were called also with BISCUIT v0.3.16.20200420 (<https://github.com/huishenlab/biscuit>), Bis-SNP v1.0.1 (Yaping Liu et al. 2012), BS-SNPer v1.1 (Gao et al. 2015) and MethylExtract v1.9.1 (Barturen et al. 2013). Default parameter settings were used, with the exception of minimum MAPQ and BQ thresholds which in all cases were set both to 1. Please refer to Suppl. Table D.1 for the complete command line in each case. The resulting variant calls were normalised, decomposed and otherwise processed for comparison to the high-quality reference data using BCFtools v1.9 (H. Li et al. 2009; Danecek et al. 2021).

6.2.4 Benchmarking

Benchmarking itself was performed with vcfeval of RTG Tools v3.11 (Cleary et al. 2015), which compares both the substitution context and estimated genotype of baseline variants from the truth set to each set of calls from bisulfite data in order to evaluate true positives, false positives and false negatives, in response to varying common filtering thresholds such as sequencing depth (DP), quality (QUAL) and genotype quality (GQ). Variants must occur with both the same substitution context and genotype in order to be evaluated as a true positive. As described in Chapter 5, sensitivity (recall) refers to the true positives as a fraction of the truth set positives, described in equation (5), whereas precision, described in equation (6), refers to the true positives as a fraction of the discovered variants (Table 6). The F1 score, described in equation (7), reflects the balance of precision and sensitivity via the harmonic mean of both measures, and can be optimised relative to each filter by taking the maximum value in response to varying the relevant threshold.

6.3 Results

In benchmark data sets for both test species, precision-sensitivity of the SNPs derived from WGBS data is demonstrably improved following double-masking in comparison to existing methods (Figure 26). Notably, common filtering metrics such as variant quality (QUAL) and genotype quality (GQ) behave as could be expected in conventional sequencing data (Figure 26; Figure 27), facilitating in many cases the use of established best-practice criteria for selecting high-confidence calls. Additional comparison of SNPs derived from real WGS data (*A. thaliana*; accession Cvi-0) and equivalent WGBS data, following *in silico* bisulfite conversion (~99%) of sequencing reads, removes the variation caused by differences during sequencing, but not alignment. The resulting

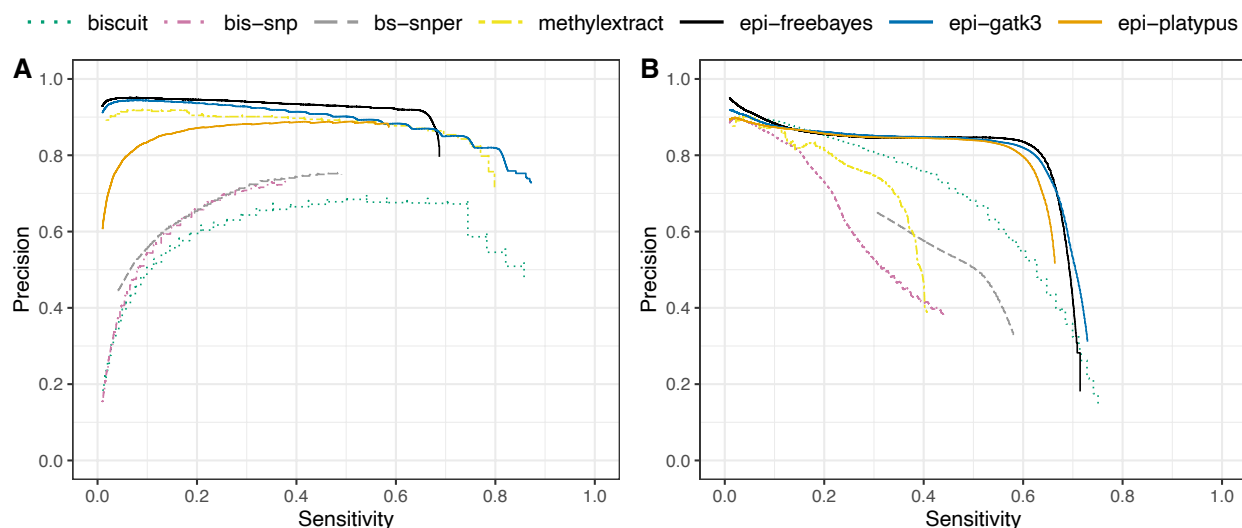


Figure 26. Precision-sensitivity plots demonstrating the response to an increasing variant quality (QUAL) threshold, comparing SNPs derived from published WGBS data to those derived from established benchmark datasets for **(A)** *A. thaliana* and **(B)** human. Software with the epi- prefix are intended for conventional sequencing libraries but in this case run after pre-processing with the double-masking procedure. True and false positives are evaluated based on both the substitution context and the estimated genotype.

ROC-like curves demonstrate a comparable level of sensitivity (i.e. true positives) in both WGS and WGBS data following variant calling with Platypus, Freebayes and GATK3.8 UnifiedGenotyper (Figure 28), however there is a drop in precision driven in each case by an influx of false positives. When *in silico* bisulfite conversion is instead applied directly to the WGS alignments, thus eliminating variation due to the alignment of bisulfite-treated reads, the differences in false positives are reduced for each tool. All software demonstrate an appreciable performance, with GATK3.8 achieving the highest raw number of both true and false positives, followed by Freebayes and then Platypus, for both WGS and WGBS data. The total number of false positives derived from *in silico* WGBS alignments however represent only 1.0%, 3.8% and 4.3% of the total, unfiltered calls for those same tools respectively, when discounting the fraction shared in the equivalent WGS data.

The overall balance between precision and sensitivity can be evaluated using the harmonic mean, to denote the F1 score, which can be compared between different software and data types (Table 7). With *in silico* WGBS reads, the optimal F1 scores for GATK3.8, Freebayes and Platypus were identified at 0.8508, 0.8039 and 0.7709, respectively, with a corresponding QUAL threshold of 80, 41 and 27. The overall best-performing tool was therefore GATK3.8, achieving 0.8685 sensitivity and 0.8338 precision at the optimal level, followed by Freebayes with 0.7335 sensitivity but a higher precision of 0.8894. Freebayes performed more similarly between the *in silico* WGBS reads and

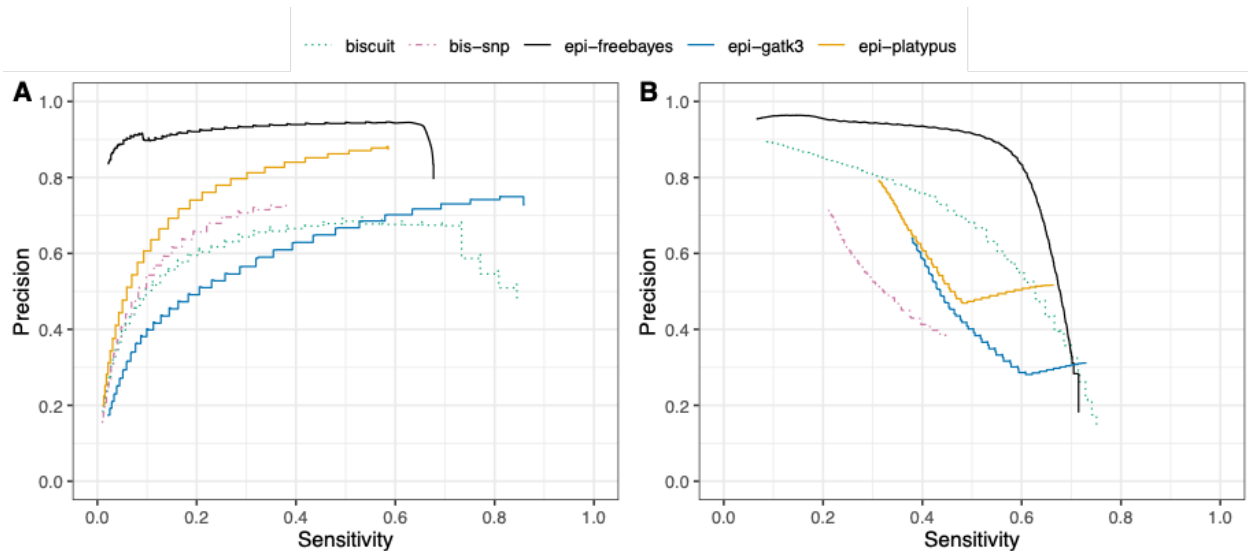


Figure 27. Precision-sensitivity plots demonstrating the response to an increasing genotype quality (GQ) threshold, comparing SNPs derived from published WGBS data to those derived from established benchmark datasets for **(A)** *A. thaliana* and **(B)** human. Software with the epi- prefix are intended for conventional sequencing libraries but in this case run after pre-processing with the double-masking procedure. True and false positives are evaluated based on both the substitution context and the estimated genotype. With GQ as qualifier, epi-freebayes performs consistently high in terms of the optimal balance of true and false positives, with an F1 score of 0.7715 and 0.6984 in each dataset, respectively. BS-SNPer and MethylExtract do not give values for genotype likelihood or genotype quality in the output VCF files. Additionally, the reported GQ values in biscuit are identical to the QUAL values in single-sample mode.

the real WGBS dataset, however, suggesting it may account better for differences in library composition and layout. Platypus performs better overall in default mode, despite an optimal precision level of 0.9436 for WGS and 0.8991 for WGBS data with assembly-mode enabled (not shown). The reduced overall performance due to lower sensitivity may in-part arise due to the need to set a pre-emptive threshold for Platypus at $BQ \geq 0$ (`--minBaseQual=0`), following the double-masking procedure, to avoid over-filtering regions during local assembly.

Table 7. Optimised F1 scores for *A. thaliana* (Cvi-0) in comparison to the reference SNPs obtained from 1001 genomes, when using real WGS and WGBS data, alongside *in silico* WGBS data derived from the WGS reads and alignments, respectively.

	Real data		<i>in silico</i>	
	WGS	WGBS	reads	alignments
GATK3.8	0.9189	0.8177	0.8508	0.9069
Freebayes	0.8247	0.7670	0.8039	0.8247
Platypus (default)	0.7423	0.7026	0.7709	0.7935
Platypus (assembly)	0.6378	0.5980	0.6449	0.6509

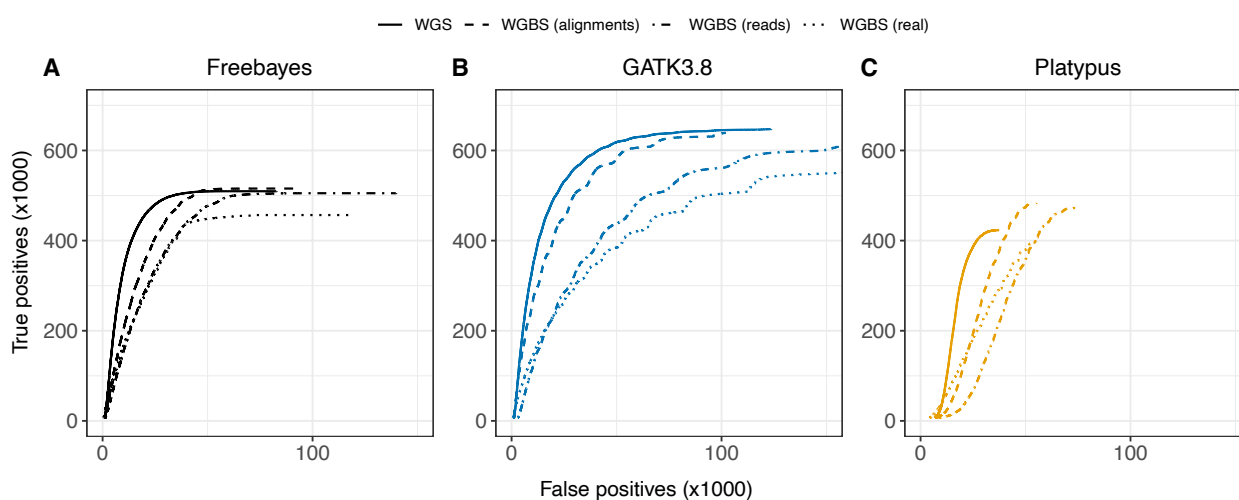


Figure 28. ROC-like plots demonstrating the response to an increasing variant quality (QUAL) threshold, comparing SNPs derived from real data (WGS) to those derived from equivalent data (WGBS) after *in silico* bisulfite conversion of either reads or alignments, followed by pre-processing with the double-masking procedure, in *A. thaliana* (Cvi-0). The real WGBS dataset from Figure 26A is also displayed alongside in each panel for comparison. Panels show results from conventional software **(A)** Freebayes, **(B)** GATK3.8 and **(C)** Platypus (default mode). True and false positives are evaluated based on both the substitution context and the estimated genotype.

When considering only those variants called by GATK3.8 UnifiedGenotyper, the relative fraction of true and false positive variants shared between each dataset, before and after filtering according to GATK best-practices (described in Supplementary Table D.2), helps to further decompose the factors mainly responsible for the differences observed with WGS and WGBS data (Figure 29). For example, among the unfiltered true positives the majority of variants are similar and shared between all datasets, with a smaller, secondary, sub-fraction shared only among the real WGS data and both simulated WGBS datasets (paired-end, ~62X). After filtering, the number of true positive variants are reduced mainly in the real WGBS dataset (single-end, ~34X), suggesting that variable sequencing library composition is driving these differences. Upon further inspection, the filter on StrandOddsRatio (SOR) appeared to be excluding the majority of true positive variants filtered out in the real WGBS data, likely as a result of an indirect strand-specificity imposed on potential variant calls by the double-masking procedure. When filtering the true positives in the same manner from the real WGBS dataset in the NA12878 human line (Figure 26B; paired-end, ~46X), however, these variants were only reduced by ~13%. With some low-coverage libraries it might therefore be prudent to relax the SOR filter when seeking to obtain confident calls from WGBS data. The false positives, on the other hand, are reflected primarily in the real WGBS dataset and the artificial dataset simulated from real WGS reads (subsequently aligned as a WGBS library). Here, it is the variant confidence metrics (i.e. QUAL and QualByDepth) which are driving the differences after

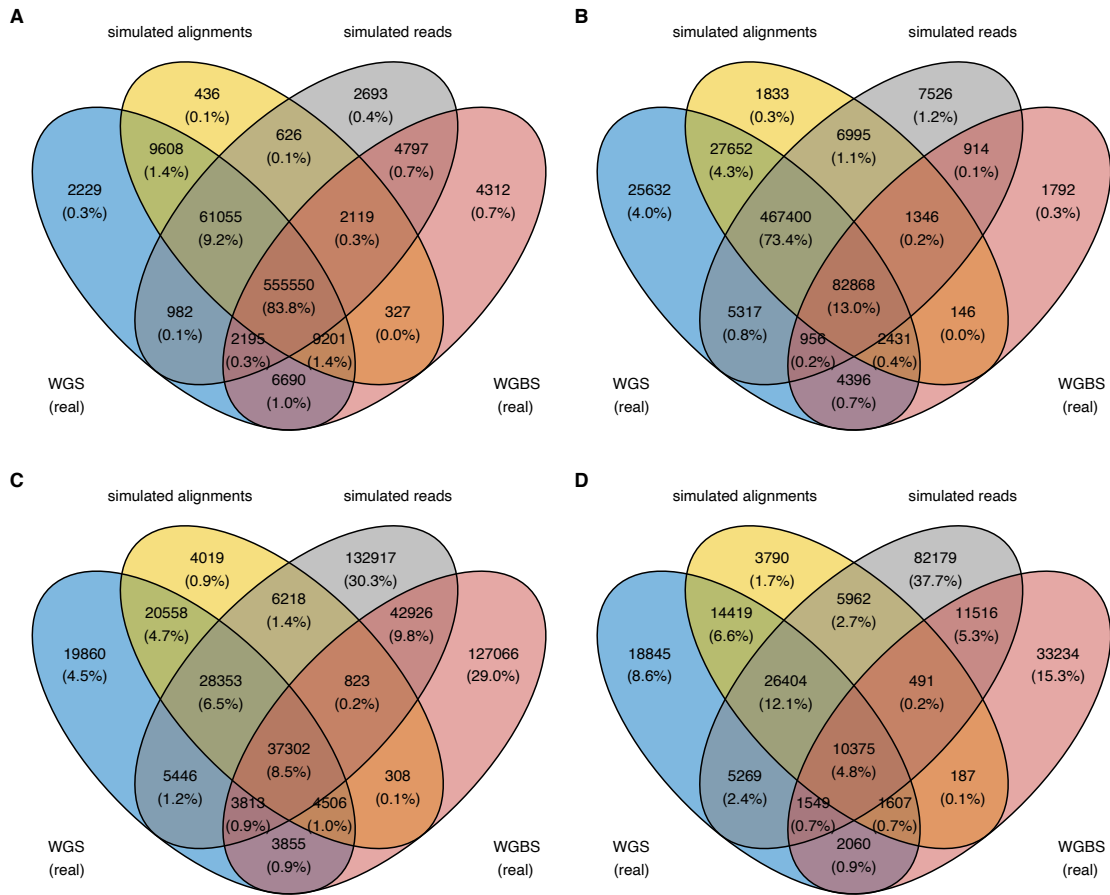


Figure 29. The shared fraction of true and false positive variants in real and simulated data for *A. thaliana* (Cvi-0), following analysis with GATK UnifiedGenotyper. Distinct WGBS datasets were simulated from both the real WGS alignments and the real WGS reads, separately. The panels denote (A) true positives, before and (B) after filtering, according to recommended hard-filter thresholds in GATK best-practices, and (C) false positives, also before and (D) after filtering. The thresholds chosen for filtering are further described in Supplementary Table D.2.

filtering. Taken together this further suggests that the influx of false positives relative to real WGS data are driven primarily by differences in both alignment and library composition, both of which have a direct influence on variant calling.

This indirect strand-specificity imposed on potential variant calls by the double-masking procedure can be expected to reduce the available sequencing depth required to make confident calls for potential polymorphisms involving thymine, in comparison to WGS data. In the equivalent, *in silico* WGBS library derived from WGS reads, this would seem to manifest predominantly as a relative decrease in variant confidence metrics on true positive SNPs (Figure 30). The number of true positive variants that would fail the recommended hard-filtering thresholds ($QUAL < 30$ or

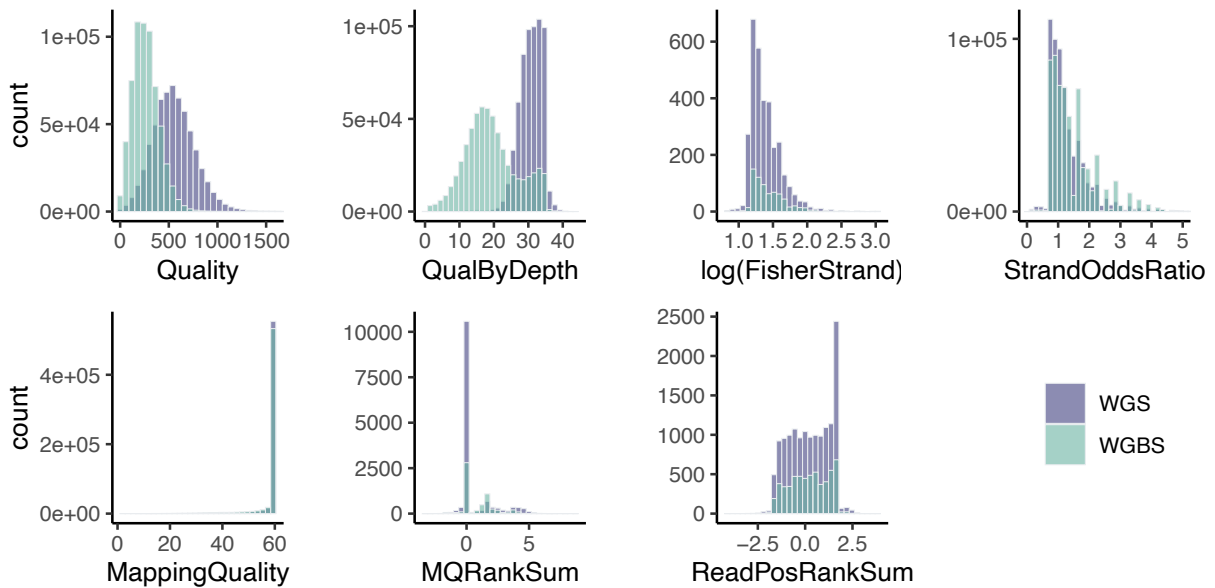


Figure 30. Quality control metric comparisons between SNPs derived from real data (WGS) and those derived from equivalent data (WGBS) after *in silico* bisulfite conversion of WGS reads, in *A. thaliana* (Cvi-0). Distributions are derived from the intersection of true positive calls made in each case by GATK3.8 UnifiedGenotyper, and each metric is considered specifically for hard-filtering in GATK best-practices. Definitions for each metric according to GATK are given in Suppl. Table D.2. MQRankSum and ReadPosRankSum are only evaluated for a subset of calls, and in both datasets the vast majority of calls achieve a FisherStrand score of 0 (not shown).

QD<2.0), however, increased only from 1,730 (<0.27%) in WGS data to 9,762 (<1.55%) in the *in silico* WGBS data. In this simulated, paired-end library there is only a minor increase in overall strand bias, as measured with the SOR metric in GATK3.8 UnifiedGenotyper, where true positive variants that would fail the recommended hard-filtering threshold (SOR>3) increased from 18,045 (2.79%) in WGS data to 31,487 (5.0%) with simulated WGBS data. All together the number of true positive variants lost after hard-filtering increased from 30,858 (4.77%) to 56,695 (9.0%) due to the *in silico* bisulfite conversion, while the total false positive variants increased from 80,528 (6.81%) to 143,745 (10.24%).

Between all selected variant callers, the proportional deviation of false positives from *in silico* WGBS reads, relative to WGS data, show similar profiles when partitioned by substitution context (Figure 31). A total of 92.3%, 77.3% and 72.8% of the total false positives here occur in positions which are homozygous-reference in the truth set for each of GATK3.8, Freebayes, and Platypus, respectively, after filtering those shared in the equivalent WGS data. These positions represent 12.0%, 5.6% and 5.6% of the total, unfiltered calls made by each tool. The remaining false positives typically comprise true variants which have been assigned an incorrect genotype (e.g. homozygous-alternative called as heterozygous), representing 2.9%, 4.2% and 4.6% of the total, unfiltered calls.

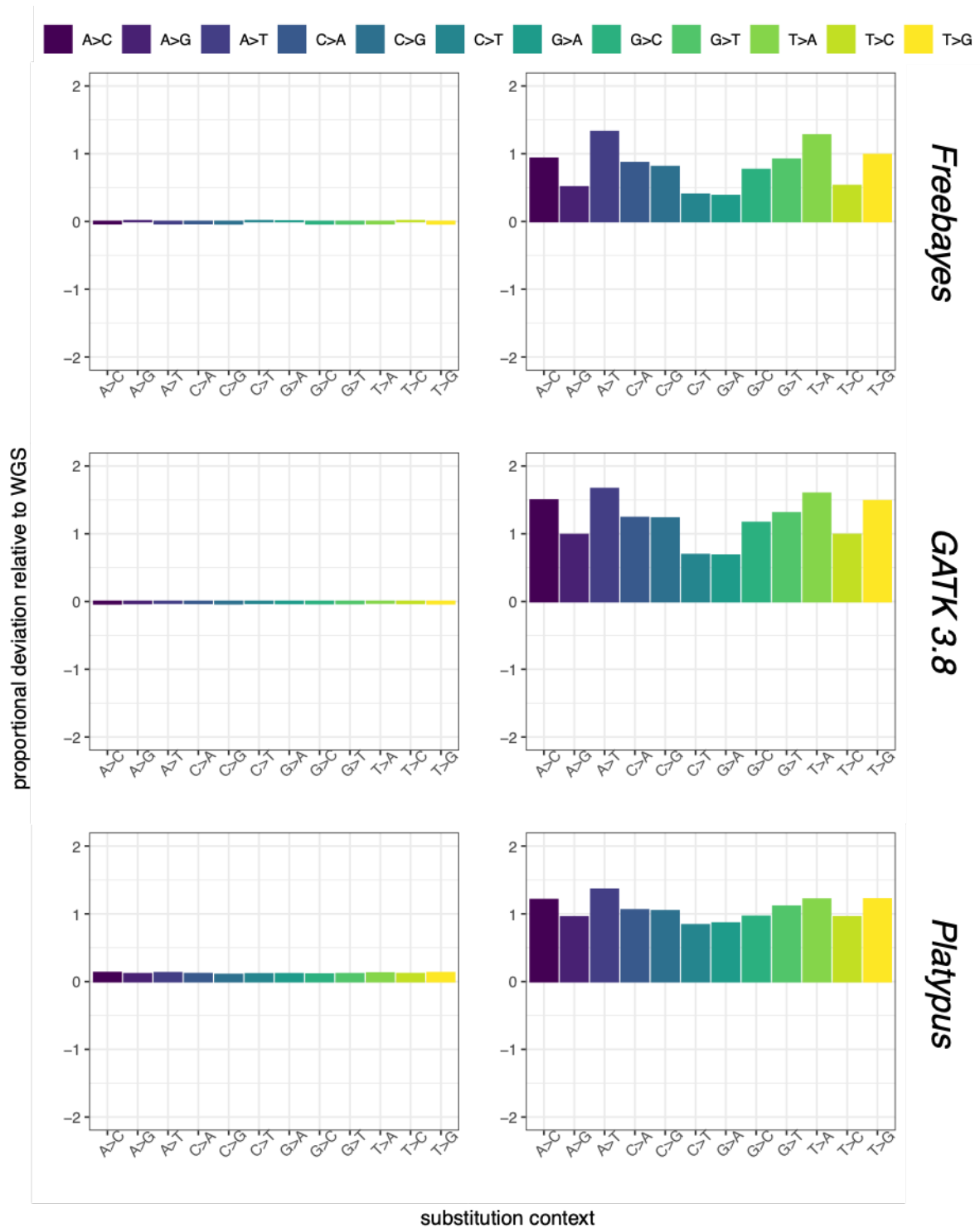


Figure 31. Proportional deviation of raw, unfiltered variant calls from data subject to *in silico* bisulfite treatment (WGBS) relative to the original untreated WGS data, in *A. thaliana* (Cvi-0). The left-hand panels indicate the fraction of true positive sin each dataset, whereas the right-hand panels refer to the fraction of false positives. The profiles partitioned across each substitution context are similar between each tool, although there are fewer false positives in Platypus (default mode) and Freebayes relatives to GATK3.8, and in contrast to the other tools the true positives called by Platypus seem to increase overall in the artificial WGBS data relative to WGS.

Many of these cases suffer a low GQ likely as a consequence of reduced sequencing depth by limiting calls in bisulfite contexts to opposite-strand alignments. Such positions are also considered among the false negatives, alongside the fraction of true SNPs which are not called at all from bisulfite data. When considering the sequencing depth distribution of false negatives from *in silico* WGBS alignments, discounting those shared in the WGS data, there is a peak at $\sim 4\text{-}5\text{x}$ in addition to a larger peak which correlates with the distribution for the true positives at $\sim 18\text{-}20\text{x}$ (not shown). Accounting for a minimum per-position sequencing depth of $\sim 7\text{-}10\text{x}$ should generally therefore be enough to make a successful call, disregarding differences due to WGBS alignment or significant deviations from typical sequencing biases (e.g. strand bias). More generally, aiming for a genome-wide coverage of at least $\sim 40\text{X}$, using a paired-end, directional library, would appear to be the optimal recommendation for analysis based on the complete results of this study.

6.4 Discussion

Conventional germline variant callers can be broadly categorised as alignment-based, such as GATK3.8 UnifiedGenotyper, or haplotype-based, such as Freebayes and Platypus. Both strategies are concerned with correctly identifying variants at a given locus and inferring probabilistic genotype likelihoods based on allelic count differences, however they differ in their consideration of proximal variants to establish phase. Whilst UnifiedGenotyper considers precise alignment information in a position-specific, independent manner, Freebayes considers the literal sequence of each overlapping read to obtain the context of local phasing and derive longer haplotypes for genotyping. Some modern variant callers, including for example Platypus and GATK HaplotypeCaller, expand upon the haplotype-based approach by incorporating local assembly to aid in resolving potential indels. Bisulfite sequencing data can be made conceptually compatible with each of these described approaches, following pre-processing with the double-masking procedure, with the caveat that the chosen software for calling variants handles base quality specifically during the estimation of genotype likelihoods, ideally with an option for hard-filtering. Local assembly presents an added difficulty in that base quality is often considered additionally for read trimming during construction of De Bruijn graphs, e.g. in determination of "ActiveRegions" in GATK HaplotypeCaller, and is typically codependent on the same parameter used for setting its threshold during Bayesian inference. This can sometimes be circumvented, as demonstrated herein with Platypus, by allowing even a base quality of zero during local assembly before relying on the genotype likelihood model to weight such positions appropriately during variant calling, but such a case is not ideal. If masked nucleotides are allowed to be included in the model for deriving

genotype likelihoods then the allelic balance on each variant will skew towards any mutations arising from bisulfite conversion, leading to a greater incidence of false positives.

To the best of my knowledge, the software chosen for comparison during this benchmark analysis represent almost the full extent of available, bisulfite-aware variant callers. In one instance a tool had to be omitted for both reasons of compatibility and because we were unable to run the variant calling aspect outside the context of a larger pipeline. gemBS (Merkel et al. 2019) is a pipeline suite which includes mapping, quality control, variant calling and extraction of methylation values. Attempts to run just the variant calling aspect (`bs_call`) using the standard alignment files generated in this study were unsuccessful, meaning we had to re-run the mapping too with gemBS, thus introducing a discrepancy in comparison to other tools. Furthermore, the variant output was returned in a custom, non-standard VCF format which made it very difficult to separate sequence variants from methylated sites in a manner which was also conducive to a fair, systematic comparison with the other variant calling software. These results were thus omitted so as not to disadvantage gemBS under an experimental design which may simply not be elaborate enough in this case for a fair and robust evaluation of its performance.

It is important to consider that, unlike most other bisulfite-aware tools, variant calling with the presented approach is almost completely dissociated from the influence of cytosine methylation. The advantage of this is an improved sensitivity for high-confidence variants with fewer false positives, whilst preserving the underlying model of selected tools, but the methylation level itself must be evaluated independently. This is akin to several variant-independent approaches such as MethylDackel (<https://github.com/dpryan79/MethylDackel>) and GATK MethylationTypeCaller which are commonly used to estimate the methylation level without knowledge of the underlying SNPs. In combination with the presented approach it would be feasible to derive accurate variant-adjusted methylation calls, or even allele-specific methylation without the need for a corresponding genotype dataset obtained by conventional DNA sequencing.

In conclusion, the double-masking procedure facilitates sensitive and accurate variant calling directly from bisulfite sequencing data using software intended for conventional DNA sequencing libraries. The procedure can be readily adapted to existing software pipelines and does not necessitate any additional understanding of customised VCF files. Given sufficient sequencing depth, accurate alignment with minimal deviation from expected sequencing biases, and an appropriate level of filtering based on variant quality metrics, the SNPs derived from WGBS data

are comparable to those from WGS data. The method presents a viable, alternative strategy to those who would otherwise need to sequence corresponding libraries of each type in order to better understand the role of DNA methylation in the context of the genetic background.

6.5 A pipeline for SNP variant analysis

Based partially on the results of this benchmarking analysis, a computational pipeline for calling SNP variants from bisulfite sequencing data was developed for use by the EpiDiverse consortium in the study of epigenetics in plant ecology. It comprises one aspect of the “EpiDiverse Toolkit”, and makes use of common file formats and standards to facilitate interoperability both with other pipelines in the toolkit and external software. The pipeline is implemented with Nextflow under the DSL v2 framework, and facilitates processing of population-scale data in a highly-parallel manner. The system is portable and able to be easily-configured for different computational architectures, with very little bioinformatic expertise required on behalf of the end-user. The software is available at <https://github.com/EpiDiverse/SNP>

The workflow first processes alignment data from SAM/BAM inputs, such as those provided as output from the EpiDiverse/WGBS pipeline, which can be provided either as a series of independent files for parallel, sample-specific variant calling, or as a single, combined file for joint variant calling, whereby each sample is represented by a read group identifier. The pipeline performs alignment sorting, indexing and recalculates proper MD tags in the input files using SAMtools (Li et al. 2009), then uses the double-masking approach to format bisulfite data appropriately for variant calling. Variant calling itself is performed using FreeBayes, and downstream post-processing and filtering performed using BCFtools (Li et al. 2009; Danecek et al. 2021).

7 Population-level Epigenomics

7.1 Introduction

The extent by which genetic diversity influences adaptation and selection, by processes of stochastic mutation, genetic drift and gene flow, is a fundamental concept underpinning modern evolutionary synthesis. Genetic diversity can be defined as any measure that quantifies the magnitude of genetic variability within a population. On the other hand, phenotypic plasticity is defined as the capacity of a set genotype to produce more than one phenotype when exposed to different environmental conditions (Kelly, Panhuis, and Stoehr 2012). Intraspecific trait variability is a direct result of phenotypic plasticity, contributing to increased functional diversity within plant communities, and is a key component of biodiversity with important implications for species coexistence and ecosystem functioning (Medrano, Herrera, and Bazaga 2014). Genetic diversity can therefore be considered the baseline for phenotypic diversity, upon which evolutionary processes like natural selection can act (Hughes et al. 2008).

In recent years, however, it has become evident that epigenetic variation can also play a role in phenotypic plasticity (Bossdorf, Richards, and Pigliucci 2008; Heer et al. 2018), and several studies have furthermore suggested that it can contribute to functional diversity in populations (Latzel et al. 2013). For example, epigenetic mechanisms play a role in allelopathy, and epigenetic changes might be more determinant than genetic variability in the success of plant invasions (Pérez et al. 2012; Hofmann 2015; Slotkin 2016). Epigenetic variation can also have a role in how plants respond to environmental stress conditions (Kinoshita and Seki 2014). Transgenerational inheritance has been demonstrated for example in *A. thaliana* through the use of “epigenetic Recombinant Inbred Lines” (epiRILs), unveiling DNA methylation as a possible source of heritable phenotypic variation whereby epialleles can influence complex traits in the absence of DNA sequence change (Johannes et al. 2009; Reinders et al. 2009). Moreover, mutation accumulation lines in *A. thaliana* have revealed that in addition to stable plant DNA methylation patterns which are able to persist over many generations, the rate of stochastic epimutation is both higher than can be explained by the (lower) rate of spontaneous genetic mutation, and far more susceptible to reverse epimutation (Becker et al. 2011). Understanding how epialleles become triggered and/or released under this apparent “transgenerational instability” will provide insight as

to whether such distinct mechanisms have broad evolutionary consequences beyond the purview of genetic variability.

In the context of plant ecology, (Richards et al. 2017) describe how current efforts in epigenetic research are directed toward an understanding of i) natural patterns of epigenetic variation, ii) the origins and drivers of this variation, and iii) its ecological and evolutionary consequences. These questions centre chiefly around the study of “epigenomics” in natural populations, the interplay between genetic and epigenetic variation, and the influence (if any) of the surrounding environment. This chapter focuses on the bioinformatic basis on how to bring such findings to light, within the scope of non-model plant species, as an extension of work from previous chapters and existing work on model species.

7.2 Challenges in Population-level Epigenomics

Although epimutations may arise spontaneously, a significant fraction of all epigenetic variation found within a population has both a genetic and environmental basis (Kawakatsu et al. 2016). Significant effort must therefore be directed towards the decomposition of variance between each of these explanatory factors, which presents a challenge both in experimental design and in practical terms. NGS will form the basis of much new understanding into natural patterns of epigenetic variation in non-model species, but it is expensive to perform and even more so for population-scale data. The sheer number of samples which must be analysed at sufficient sequencing depth, in order to generate enough statistical power in population experiments, can also be a major bottleneck with limited computational resources. Leveraging the workflow management system Nextflow allows for high parallelisation and optimal usage of resources, in a framework which is both portable and easily configurable for different computational architectures (see Chapter 5).

Furthermore, while bisulfite sequencing can provide insight into the DNA methylome in order to build a picture of epigenetic variation, the chemical treatment also artificially obscures natural mutations in cytosine contexts, making it difficult to draw reliable inferences about the genetic background. To have to build corresponding sequencing libraries both for conventional and bisulfite data compounds the issue yet further when it comes to resource expenditure. The development of a new algorithm to facilitate variant calling from bisulfite sequencing data in order to reliably infer genotypes (see Chapter 6) can help considerably in bringing down sequencing costs and making population-level analyses more feasible.

In addition, downstream analyses of population-level data bring their own challenges. Much new understanding is driven by the application of methods which were devised previously for genetic data. Genome-wide association studies (GWAS) for example make use of linear models to identify SNPs within a large cohort which are associated with a quantitative trait (e.g. plant height, flowering time). Conceptually, it is valid to use epigenetic marks such as methylated cytosines in place of SNPs in such an analysis. Practically, however, it is rarely feasible to do so, due to i) the extent of background “noise” driven by highly dynamic, stochastic epimutations, ii) a tendency for epigenetic marks to convey only weak associations relative to SNPs as a result of their lower stability, and iii) the very high number of comparisons which are often necessitated in the statistical analysis of such datasets.

Whilst SNPs can have strong effects on trait variability, this does not seem to be the case for DNA methylation variable positions (MVPs), which often affect transcription only after accumulating over a broader genomic region (Cubas, Vincent, and Coen 1999; Manning et al. 2006; Martin et al. 2009). For this reason, the study of differentially methylated regions (DMRs) became very popular in high-resolution studies (Schmitz et al. 2013; Cortijo et al. 2014; Dubin et al. 2015; Kawakatsu et al. 2016). Indeed, given the complex dynamics of DNA methylation variability, it can be quantified at several different levels, from global (or genome-wide) methylation, to average methylation limited to sequence contexts or genomic features, to the methylation of specific genomic regions or individual positions (Schultz, Schmitz, and Ecker 2012). Each may be appropriate for population-scale data under different experimental circumstances.

7.3 Differential Methylation

DNA methylation is highly dynamic, and there is substantial variation at both the individual and population-level. Given a set of biological replicates, the challenge is to distinguish specific, relevant differences from stochastic epimutations among the background variation. On a genome-wide scale, this typically requires a priori knowledge of some level of sample stratification in terms of both sequence context and grouping, for example by phenotypic or environmental differences. Most experimental studies contrast DNA methylation levels from different samples to each other in this way, be it mutant background and wild type, treatment and control, or different natural accessions obtained across a broad geographical area. Statistical comparisons between samples aim to identify DNA methylation differences at either the single nucleotide level (differentially

methyated positions; DMPs) or the region level (differentially methylated regions; DMRs). While DMPs provide useful information on the rate at which epigenetic changes occur (Becker et al. 2011; Schmitz et al. 2011; Van Der Graaf and Wardenaar 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.

Given the biological relevance of DMPs and DMRs, the task then in bioinformatics is how to define them. Classification at the single nucleotide level would appear simple enough. Given a set of samples belonging to group A and group B, a Mann-Whitney U test can for example be applied over each position where the methylation rate in each case is given as a proportion of reads which are methylated. In practical terms, repeating this for every position genome-wide creates a heavy multiple testing burden, however, necessitating a post-hoc correction procedure which often leaves few significant results despite seemingly clear differences. On a region-level, the definition of what constitutes a DMR is more ambiguous. Some strategies, such as DSS (Feng, Conneely, and Wu 2014; Feng et al. 2014; Ziller et al. 2013; Schultz et al. 2016), define DMRs as clusters of spatially adjacent DMPs. These approaches are in turn subject to the same burden of multiple testing, and can struggle to return significant results. This issue is also prevalent in tools which attempt to gauge DMRs using a window- or sliding-window-based approach, as implemented, for example, in methylKit (Akalin et al. 2012). Other strategies attempt to mitigate the multiple testing problem by calling DMRs only in pre-defined regions, including e.g. gene promoters or other annotations, but these risk excluding other biologically relevant loci from the analysis.

More recent tools for DMR calling have focused on the development of alternative methods for segmentation, i.e. the pre-selection of genomic regions which are biologically relevant. Among these are metilene (Jühling et al. 2016), dmrseq (Korthauer et al. 2018) and HOME (Srivastava et al. 2019), which implement multi-step strategies that first restrict the testable genome space to candidate regions with evidence of methylation, prior to assessing statistically significant differences. Though metilene and HOME are both applicable to non-CG methylation contexts, all three of these tools have been developed with mammalian (e.g. human) DNA methylation data in mind. By extension, they are frequently built on assumptions from such data, which might include for example most cytosines being methylated, strong local correlation between methylation states, predominant (or exclusive) CG methylation, and largely binary methylation states. These often do not reflect the distinct molecular mechanisms governing cytosine methylation in plants, particularly in non-CG contexts, and indiscriminate application of such assumptions to plant data might result

in many methylated cytosines and regions being falsely discarded. It is prudent to assess on a case-by-case basis which tool is most suitable based on whether their assumptions hold regarding the specific research question to be addressed. In the absence of a universal, biological definition of a DMR, however, building on the findings of such tools represents the current best-practice.

7.3.1 A pipeline for case/control DMRs

Under the broader initiative of the EpiDiverse project (<https://epidiverse.eu/>), metilene was chosen to be adapted into a workflow management system, in order to extend its functionality in the context of plant ecology. The underlying model comprises a binary segmentation algorithm combined with a two-dimensional Kolmogorov–Smirnov test, which allows for the detection of DMRs based on both spatial and methylation differences simultaneously between two groups of samples. The non-parametric nature of the statistical test makes no assumptions about the underlying distribution of the data. The software analyses whichever 5mC sites are provided in the input files, itself unaware of methylation contexts, thus allowing the end-user to provide whichever context is most relevant to them. It was considered the best choice for case/control studies with non-model plant data out of available software (Kreutz et al. 2020). Within EpiDiverse, the tool is implemented in a Nextflow pipeline under the DSL v2 framework, and facilitates processing of population-scale data in a highly-parallel manner. It comprises one aspect of the “EpiDiverse Toolkit”, and makes use of common file formats and standards to facilitate interoperability both with other pipelines in the toolkit and external software. The system is portable and able to be easily-configured for different computational architectures, with very little bioinformatic expertise required on behalf of the end-user. The software is available at <https://github.com/EpiDiverse/DMR>

The workflow first processes methylation data from BED inputs in any context, such as those provided as output from the EpiDiverse/WGBS pipeline, using BEDtools unionbedg (Quinlan and Hall 2010) to produce a combined matrix. The pipeline then performs annotation of DMRs with metilene, between any groups defined by the user in a provided sample sheet. Downstream visualisation of results is carried out using R packages ggplot2 (Wickham 2011) and gplots (Warnes et al. 2009).

7.3.2 A pipeline for population-level DMRs

All current methods for annotating DMRs based on methylation data from a set of samples until now depend on a priori stratification of samples into distinct groups for comparison. While this “forward epigenetics” approach aids greatly in controlling variance in cases where clear-cut groups can be defined, often it is not so apparent, particularly in data obtained from natural populations. It can also be the case in such multi-dimensional data that groups differ from locus to locus throughout the genome, for example based on interacting environmental factors or complex phenotypes. In the context of plant ecology, an unsupervised approach to DMR calling would aid in identifying variable regions which are responsive under certain conditions, allowing researchers to further investigate which conditions and phenotypes are explained on a basis which could be considered more akin to “reverse epigenetics”.

In collaboration with the EpiDiverse project (<https://epidiverse.eu/>) and Computomics GmbH (Tübingen, Germany), the software MethylScore was developed into a pipeline for the robust identification of DMRs from plant WGBS data, taking into account the complexity and variability of plant DNA methylation while using an informed, restricted set of candidate regions for statistical testing. MethylScore aims to avoid the necessity for pre-defining sample groups, required by existing tools, to increase its applicability to sample populations with inherent group structure and to prevent experimenter bias. The pipeline is implemented with Nextflow under the DSL v2 framework, and facilitates processing of population-scale data in a highly-parallel manner. The system is portable and able to be easily-configured for different computational architectures, with very little bioinformatic expertise required on behalf of the end-user. The software is available at <https://github.com/Computomics/MethylScore>

The differential methylation analysis module of MethylScore is built around a two-state Hidden Markov Model (HMM), as per the approach described by (Molaro et al. 2011). To identify and segment methylated regions in plant genomes, independent of prior information, the original implementation was extended beyond the CG sequence context, allowing the algorithm to train distinct parameter sets for each methylation context relevant to plant ecology. In each case, parameters are estimated for a beta-binomial distribution, accounting for both stochastic variance in the coverage distribution (assumed to be beta distributed) as well as between-sample biological variance (binomially distributed). MethylScore trains on the provided WGBS data itself, thereby forming data-specific assumptions regarding the underlying distribution of methylation patterns by

which to inform later statistical comparisons. Regions of interest are also inferred during this step, which helps to mitigate the multiple-testing burden by limiting the number of tests to such loci.

Using publicly available datasets for *A. thaliana* (1001 Genomes Consortium et al. 2016; Kawakatsu et al. 2016; Tedeschi et al. 2019; Ning et al. 2020; Y. Zhang et al. 2021; Wibowo et al. 2018) and rice (Stroud et al. 2013), MethylScore was 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. In comparison to the DMRs identified in the *A. thaliana* 1001 Genomes and Epigenomes datasets (1001 Genomes Consortium et al. 2016; Kawakatsu et al. 2016), MethylScore was able to detect known and unknown genotype-epigenotype associations.

7.4 Epigenome-wide Association Studies (EWAS)

Interest in understanding the connection between genetic diversity and intraspecific trait variability led to the development of genome-wide association studies (GWAS), which make use of linear models in population-scale data to reveal significant genotype-phenotype associations. In recent years, it has become an important tool to detect variants involved for example in complex human diseases (Bush and Moore 2012), aiding greatly in the development of new treatments, ranging from type 2 diabetes to schizophrenia (Visscher et al. 2017). In plants, GWAS has been used to uncover the genetic basis of important traits in agriculture and to accelerate breeding programs (Tibbs Cortes, Zhang, and Yu 2021), having been applied successfully in cereals such as maize (Thornsberry et al. 2001; Buckler et al. 2009), wheat (Neumann et al. 2011; Schulthess et al. 2017), and rice (Huang et al. 2010; Li et al. 2017), and a number of other crop species including soybean (Hwang et al. 2014), tomato (Lin et al. 2014), and cotton (Du et al. 2018). It has also been used to study plant response to changing environments for example in the model plant species *A. thaliana* (Atwell et al. 2010). Moreover, GWAS has revealed genomic regions related to physiological, agronomic, and fitness traits such as plant height, stress tolerance, flowering time, kernel number, and grain yield, and identified genes connected with geographical deviation and local adaptation.

Despite the advances made using GWAS, there remains a substantial proportion of unexplained causality which might be driven for example by epigenetic mechanisms, leading to the development of epigenome-wide association studies (EWAS) as a counterpart to the genetic approach (Rakyan

et al. 2011). Whilst conceivable that epigenetic markers can be analysed in such a manner, this approach yet brings its own array of challenges. First and foremost, DNA alleles do not typically vary across cells and with modern approaches can be genotyped with low error rates. By contrast, methylation states may be tissue-specific, and are often more akin to somatic variants at the single cell level. They can vary over alleles within a cell and, in rare cases, over DNA strands within an allele (hemi-methylation). The methylation state measured at a 5mC position is given as a proportion of total reads, reflecting the average over cells, alleles and strands, and is further obfuscated by measurement error. Genetic markers are both stable and not confounded by reverse causality, resulting in typically stronger associations than epigenetic equivalents, which can be inherited in the germline, environmentally induced, arise spontaneously, or as a consequence of disease, for example. Epigenetic changes are thus more dynamic, making it difficult to discern a significant relationship between phenotype and epigenetic mechanisms - a major challenge of EWAS (Paul and Beck 2014). Furthermore, GWAS is already statistically confounded by the extent of multiple comparisons (Pe'er et al. 2008). This issue is exacerbated yet further with epigenetic marks such as 5mC, as potential sites are typically far more prevalent genome-wide than SNPs. Data which is missing in some cohorts due to e.g. differences in sequencing depth poses yet another challenge for some EWAS methods (Pan et al. 2016).

In plants, transgenerational epigenetic marks can be transmitted to descendants both through mitosis (in case of vegetative propagation) or meiosis (sexual reproduction) (Heard and Martienssen 2014). Natural variation may therefore give rise to “epialleles”, leading to phenotypic changes that are heritable, as shown for example in *A. thaliana* where certain stress-induced transgenerational reactions depend on DNA methylation (Boyko et al. 2010; Lang-Mladek et al. 2010). In the pursuit of understanding how heritable epialleles influence plant evolution, phenotypic traits, and fitness, EWAS may be a powerful method to reveal epigenetic marks associated with biological traits (Rakyan et al. 2011; Lappalainen and Greally 2017). Thus far, however, there has been very scarce use of EWAS for plants (for example, a PubMed search for “ewas plant” returned seven hits as of 19 February 2021, while “ewas human” returned 131). Published examples include DNA methylation variation in *Quercus lobata* (valley oak) associated with climatic gradients (Gugger et al. 2016), and EWAS has been successfully applied to identify the epigenetic change that causes the metastable somaclonal variant in *E. guineensis* (oil palm) (Ong-Abdullah et al. 2015). Another study with *Pinus pinea* (stone pine) showed that there was a remarkable level of phenotypic plasticity. Vegetatively propagated *P. pinea* trees showed a high degree of DNA methylation under different environmental conditions (Sáez-Laguna et al. 2014).

7.4.1 A pipeline for EWAS analysis

Available methods to perform EWAS remain scarce, particularly in the context of plant ecology. Some software are specific for human studies such as GLINT (Rahmani et al. 2017), or unable to deal appropriately with missing data such as EWAS: epigenome-wide association study software v2.0 (Xu et al. 2018). The R package Gene, Environment and Methylation (GEM) (Pan et al. 2016), adapts an existing package for genetic data, MatrixEQTL (Shabalin 2012), in order to implement linear regression models focusing on genetic variants, epigenetic marks, and the interaction between them in a computationally efficient manner. The GEM package was herein further adapted to a workflow management system, under the broader initiative of the EpiDiverse project (<https://epidiverse.eu/>), in order to extend its functionality in the context of plant ecology. This pipeline comprises one aspect of the “EpiDiverse Toolkit”, and makes use of common file formats and standards to facilitate interoperability both with other pipelines in the toolkit and external software. The pipeline is implemented with Nextflow under the DSL v2 framework, and facilitates processing of population-scale data in a highly-parallel manner. The system is portable and able to be easily-configured for different computational architectures, with very little bioinformatic expertise required on behalf of the end-user. The software is available at <https://github.com/EpiDiverse/EWAS>

7.3.1.1 Missing data estimation

Aside from improved usability and computational advantages brought about by the workflow management system Nextflow, notable among the expanded features of the pipeline is the handling of missing data and subsequent imputation of methylation values. Often methylation values cannot be recovered in certain loci due to technical limitations, such as normal variation in sequencing depth. In such (non-random) cases, calculating the association from only the methylation values of the remaining samples can produce a bias that skews the interpretation of results. Typically the way to handle this in linear models is to avoid the test (i.e. of a given 5mC position) entirely, or to provide some kind of missing data imputation which preserves the underlying data structure of the remaining samples while minimising the effect on the model of samples with missing data.

GEM replaces missing methylation values by calculating the global methylation of the sample, and does not discard any positions with a high amount of missing data. This is conceptually similar to imputation in genetic data, whereby proximal SNPs can be in linkage disequilibrium with each other and thus have a greater likelihood of co-occurrence. In methylation data, however, the

sample-specific global methylation profile can be vastly different to the methylation level of a given 5mC position, which may contribute to a significant effect relative to the level of other samples in the cohort. The new pipeline instead adapts a similar method suggested by the software metilene (Jühling et al. 2016), a tool for calculating DMRs, which presents a similar situation in regards to missing data. Missing methylation values are thus imputed instead based on randomly sampling a beta-distribution, derived from the methylation values of the remaining samples at the same 5mC position (Raineri, Dabad, and Heath 2014). Estimated data should therefore have the minimum possible impact on the model, and any significant associations that arise should therefore be driven by the samples for which real data is available.

7.3.1.2 Mitigating the effect of multiple comparisons

Due to the high number of possible 5mC positions genome-wide, there can often be hundreds of thousands of independent tests during EWAS which need subsequent correcting for type II statistical error. Both Bonferroni and Benjamini-Hochberg correction procedures can be applicable under the given circumstances, but often make it infeasible to discern anything of biological significance under practical experimental conditions. Reducing the number of tests based on a priori information is otherwise the most direct way to boost statistical power, but can be difficult to reconcile when the purpose of EWAS is usually to identify new, candidate markers which are associated with a phenotypic trait. One approach would be to subset the data based on a set of known genes or genomic elements which are thought to be influenced somehow; another approach is to subset the number of tests based on the data itself. A typical procedure for example would be to remove positions which show zero deviation in methylation level across all samples. Alternatively, an option would be to subset the data based on a high-confidence set of MVPs, i.e. 5mC sites which are known to be more variable than others on a population-level.

As the biological consequences of methylation variation are frequently realised following the accumulation of MVPs on a region-level, DMRs are often characterised in high-resolution studies in order to provide candidates for further investigation of nearby genomic elements. Such regions may also prove useful in EWAS experiments, as they might be expected to convey stronger and more stable methylation patterns than individual MVPs. With this in mind, the feature to provide DMRs as markers in an EWAS experiment was implemented into the EpiDiverse pipeline. To exemplify this, a small analysis was performed based on methylated sites in CG context from a wild population of *Populus nigra* cv. 'Italica', which were subset according to significant DMRs discovered in the same context, as obtained from the EpiDiverse WGBS and DMR pipelines,

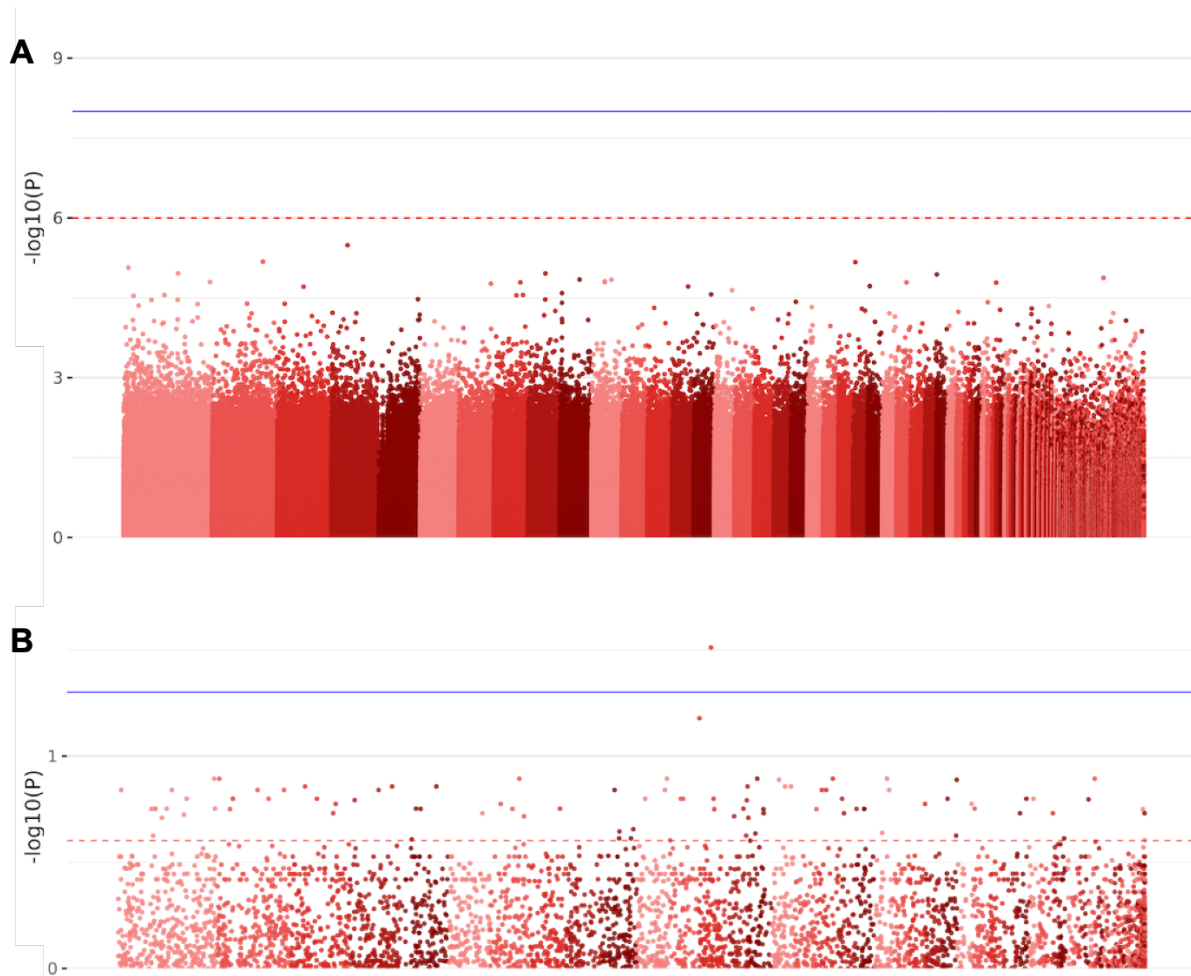


Figure 32. Manhattan plots demonstrating **(A)** the total number of tested positions during EWAS, from a cohort of *P. nigra* samples obtained from populations in Germany and Lithuania, and **(B)** the same analysis performed using significant DMRs instead. At the position-level, none were found to be significant ($p < 1 \times 10^{-8}$) or even suggestive ($p < 1 \times 10^{-6}$) based on common thresholds selected to account for the burden of multiple testing. At the region-level it becomes feasible to use Benjamini-Hochberg adjusted p -values (q), where 92 tests were found below a significance threshold of $q < 0.25$ and one even at $q < 0.05$.

respectively. EWAS was performed using leaf flavonol content measured in the parent generation as a phenotypic trait. The resulting Manhattan plot in Figure 32A reveals initially no significant QTLs below the common significance threshold of $p < 1 \times 10^{-8}$, or even below the suggestive significance threshold of $p < 1 \times 10^{-6}$, based on a global analysis of all methylated sites. The same analysis when conducted at the region-level, however, revealed a total of 92 significant QTLs ($q < 0.25$) which could be taken forward for further investigation (Figure 32B). A brief inspection of these regions intersected with functional annotations from the *P. nigra* genome returned some features potentially relevant to flavonol content, including genes with homology to ascorbate-specific transmembrane electron transporter 1, caspase family protein and mechanosensitive ion channel protein 3 alongside also methyltransferases PMT2/PMT24.

7.5 Genotyping-by-Sequencing (epiGBS)

Most bioinformatics tools and workflows discussed until now have centred on the best-case scenario in terms of the overall level of data availability to study DNA methylation, i.e. whole genome bisulfite sequencing, in a species which has an appropriate reference genome for applications such as short-read mapping and methylation calling. When studying DNA methylation in non-model plant species, however, often such a best-case scenario is not appropriate. Such experiments in plant ecology frequently have to deal with the absence of a reference genome assembly, and a high practical cost of sequencing at appropriate depth on a population-level, which is yet further compounded by the need to accommodate a simultaneous comparison of both genetic and epigenetic data, for instance to examine how much of the overall epigenetic variation between samples can be predicted from pairwise genetic relatedness (Richards et al. 2017).

A less comprehensive but cheaper and versatile alternative to WGBS is to perform bisulfite sequencing in reduced representations of the genome, for example by using restriction enzyme fragmentation during the library preparation (e.g. RRBS (Meissner et al. 2005), epiGBS (van Gurp et al. 2016), BsRADseq (Trucchi et al. 2016), epiRADseq (Schield et al. 2016) and Creepi (Werner et al. 2020)). Notable among such techniques is epiGBS, a reduced representation DNA methylation analysis tool which combines both cytosine-specific quantitative DNA methylation levels and variant calling from the same bisulfite-converted samples, with the *de novo* reconstruction of consensus reference sequences of the targeted genomic loci. This means that the method can be applied also when no reference genome is available for the species under study (van Gurp et al. 2016), albeit at a reduced level of accuracy. A similar approach was also described by (Werner et al. 2020), who provide a proof-of-concept via data obtained from almond and a PstI single enzyme digest.

7.5.1 Extending the epiGBS pipeline

In collaboration with the EpiDiverse project (<https://epidiverse.eu>), the epiGBS pipeline was improved and expanded to epiGBS2, which consists of both a detailed, updated laboratory protocol and a revised computational analysis pipeline that is accessible for all with basic knowledge in bioinformatics. The computational pipeline was implemented using the workflow management system Snakemake (Köster and Rahmann 2012), which facilitates processing of population-scale data in a highly-parallel manner. Executing epiGBS2 is cost- and time-efficient and is designed for

user-friendly, reproducible and flexible analysis, allowing for an effective determination of methylation and SNP variants in a broad range of species, including plants and vertebrates such as birds (Sepers et al. 2019). The software is available at <https://github.com/nioo-knaw/epiGBS2>

One of the new features of the expanded epiGBS2 pipeline was the inclusion of the double-masking procedure (see Chapter 6) to facilitate variant calling from bisulfite-sequencing data. As the method was developed principally for WGBS data, it was therefore important to benchmark its efficiency in its application to RRBS data. The benchmarking procedure was performed as per the methods described in Chapter 6, but for a few deviations. Primarily, the *A. thaliana* accessions Col-0, Gu-0, Ler-0, C24, Ei-2 and Cvi-0 were cultivated under the laboratory conditions described in (Gawehns et al. 2022), followed by DNA extraction and reduced representation bisulfite sequencing according to the epiGBS2 protocol. The SNPs derived from these datasets were compared with publicly available variants obtained from the (1001 Genomes Consortium et al. 2016). In order to account for the reduced fraction of variants covered after sequencing by epiGBS2, the reference datasets (obtained using WGBS) were first subset according to positions in the genome with greater than zero coverage under epiGBS2. Due to the prevalence of strand bias in RRBS methods during sequencing, and because potential SNPs in methylation contexts can only be evaluated with the double-masking procedure when reads are also available from the opposite strand, a subset of true SNPs are plausibly undetectable by the epiGBS2 pipeline. To gain insight into the magnitude of this problem, additional filtering of both the epiGBS2 and 1001 genome data was performed to keep only those positions for which at least one read from each strand (top and bottom) were present in the epiGBS2 data. Owing to the homozygous diploid nature of wild *A. thaliana* accessions, heterozygous genotypes were not called in the baseline set (1001 Genomes Consortium et al. 2016). The `--squash-ploidy` option of RTG tools `vcfeval` was used to further assess variants by treating the genome as haploid, thus consolidating heterozygous and homozygous alternative calls during benchmarking.

At high SNP quality thresholds, more than 90% of the SNPs identified in the epiGBS2 reference branch corresponded with the baseline data set (Figure 33). At lower quality values a larger total of true baseline SNPs were detected by epiGBS2, however this also resulted in an increased false positive rate (for instance, only half of the called SNPs are true positives at a quality value of 10). When excluding sites that were not covered at both strands from the epiGBS2-baseline comparison (the filtered data set), sensitivity increased; approximately 80% of the remaining SNPs in the baseline set were detected by epiGBS2 (Figure 33). When including the evaluation of exact

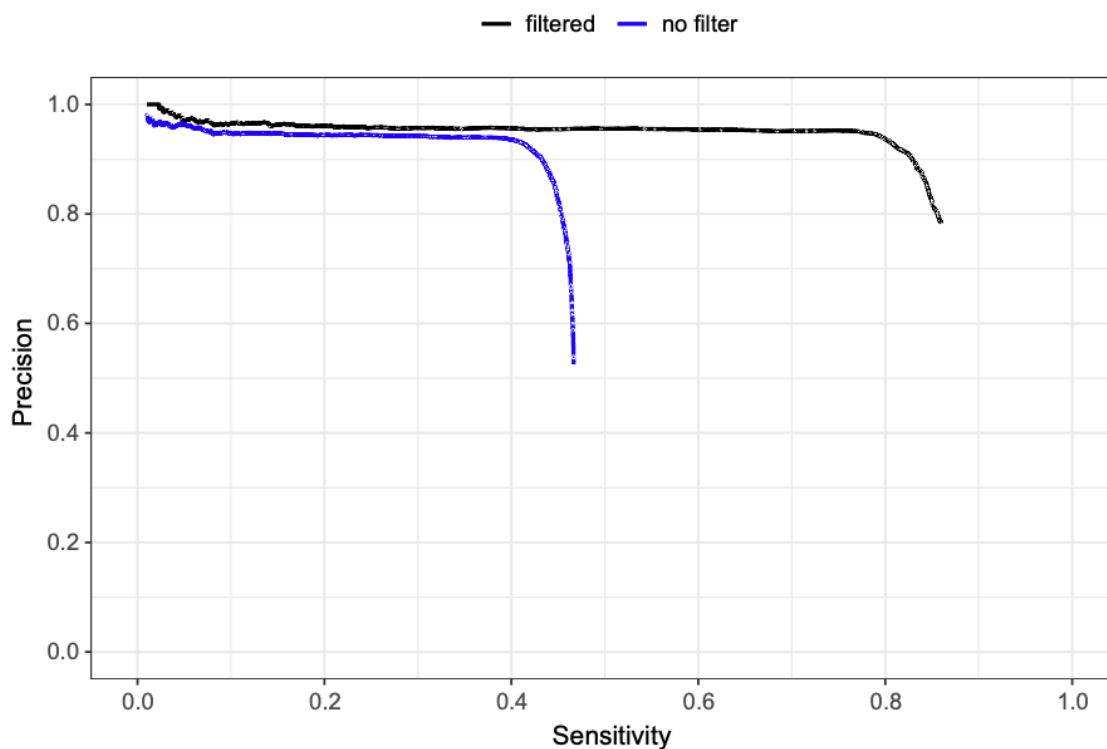


Figure 33. Precision-sensitivity of SNPs in *A. thaliana* (Cvi-0) derived using the double-masking procedure on a reduced representation bisulfite sequencing dataset, as implemented in epiGBS2. Due to the strand bias inherent to RRBS data, the evaluation was performed on both the unfiltered dataset and also a filtered subset which considered only positions from the dataset with a sequencing depth of at least 1 read on both strands.

genotype during the benchmarking, in addition to the correct allele, the overall level of precision was reduced in both filtered and unfiltered subsets. Further investigation revealed this difference to be driven almost exclusively by a fraction of true homozygous SNPs in the baseline data which were misidentified as heterozygous under epiGBS2. In contrast to the variant data from 1001 genomes, a similar excess of heterozygous SNP calls was observed in both the untreated WGS data and a corresponding set of simulated WGBS data, in *A. thaliana* Cvi-0, when using Freebayes to perform variant calling on the equivalent sequencing data. This suggests that the parameterisation of the Freebayes tool, utilised as variant caller in epiGBS2, is responsible for the small discrepancy with the benchmark data moreso than it being a direct artefact of the bisulfite sequencing.

7.6 Population-level Haplotypes

Further to the study of population-level data under the purview of plant evolution, for example in linking natural variation and intraspecific trait variability, such data can also be used to provide a statistical basis for inferring haplotypes. Often when considering sequencing data in the context of

genome assembly or later downstream analysis, little consideration is given to the variation of alleles that might be present even within a single individual. A diploid organism for example carries two copies of each chromosome, with the possibility of different sequences at the corresponding locus in each case. Yet, only one of these variants will be reflected in a typical genome assembly, which from one locus to another most likely represents an amalgamation of alleles depending on the consensus that was derived based on e.g. sequencing coverage.

Only usually during variant calling does this level of allele-specificity start to come under consideration, for example in the estimation of genotypes wherein the likelihood of homo- or heterozygosity is inferred based on the balance of alleles represented in the corresponding sequencing data. Given two proximal, heterozygous variants, however, the idea that a specific allele from one corresponds with a specific allele of the other, as would be the case given a diploid set of chromosomes, can only be addressed on the basis of “phasing”. Establishing which variants are in phase thereby allows for the reconstruction of specific haplotypes. Such haplotypes can be used to infer levels of allele-specific methylation (ASM), which has been shown to be both a hallmark of imprinted genes and also prevalent throughout the larger non-imprinted fraction of the genome (Kerkel et al. 2008). In the context of plant ecology, ASM has been investigated for example in regards to gene imprinting in maize (Zhang et al. 2014) and in understanding the RdDM pathway in *A. thaliana* (Zhang et al. 2016).

Methods for constructing haplotypes from genetic data are typically read-based, such as in WhatsHap (M. Martin et al. 2016), population-based, as in SHAPE-IT (Delaneau, Coulonges, and Zagury 2008), or a combination of both (Delaneau et al. 2013, 2019). Read-based approaches infer phase from the information provided by sequencing reads, wherein each read represents a DNA fragment which must have arisen from a single chromosomal point of origin. Variant alleles present on the same read must therefore be linked and provide evidence for local phasing. The downside of this method is that short-read sequencing approaches seldom contain two or more variants on the same read, and the technique is thus more appropriate for long-read sequencing approaches. Population-based approaches on the other hand make statistical inferences of haplotypes which capture the linkage disequilibrium (LD) from a large cohort of samples, whereby more distant variants can be resolved (Delaneau, Coulonges, and Zagury 2008; Browning and Browning 2011). The disadvantage of this approach is the large number of samples which are required in order to make inferences with enough statistical power, and it is more subject to error in comparison to the more clear-cut cases that can be observed on sequencing reads (Delaneau et al. 2013).

More recently, researchers have also begun to consider more stable epigenetic marks, such as 5mC in CG context, to provide additional context on a read-based approach (Zhou et al. 2020). Such positions are not as stable or reliable as SNPs, but occur much more frequently throughout the genome and can thus expand the number of informative reads by which to construct longer haplotypes. Though methylation level on a 5mC position is often given as a proportion of reads containing a methylated cytosine from the total number of reads, in the context of a single read the position is either methylated (cytosine) or not (thymine). The same principle regarding the chromosomal point of origin of each read, wherein each read is effectively a “mini haplotype” (Delaneau et al. 2013), must extend also to the methylation status. Each read therefore reflects the methylation status on a position-by-position basis from the same chromosome.

7.6.1 Extending the EpiDiverse/SNP pipeline

An advantage of the double-masking procedure implemented in the EpiDiverse/SNP pipeline is that methylation level can be estimated from the exact library of sequencing reads which are also used to derive SNP variants. Future work on the pipeline could involve leveraging this information in regard to both haplotype construction and deriving allele-specific methylation in turn from the same source of data. An example overview of such a pipeline is denoted in Figure 34, whereby

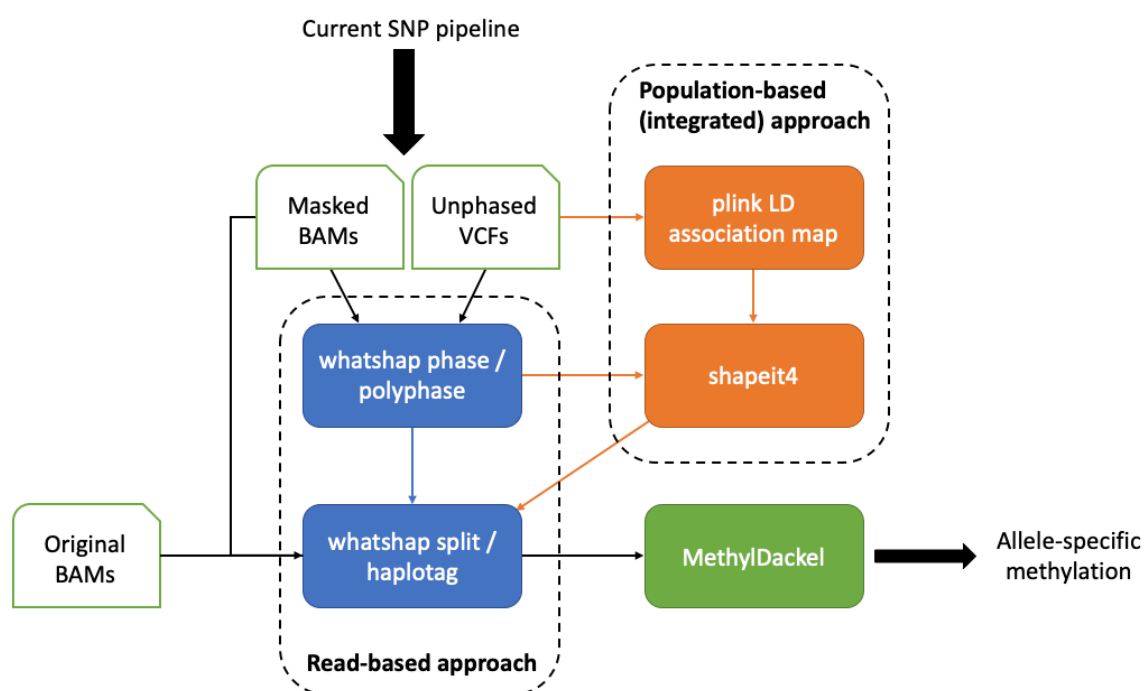


Figure 34. An overview of a proof-of-concept approach for deriving allele-specific methylation from a bisulfite sequencing library based on variants linked by reads and by analysis of linkage disequilibrium in a population-level dataset.

7 Population-level Epigenomics

double-masked alignment (BAM) files and unphased VCF files could be analysed with existing tools WhatsHap, Plink and SHAPEIT4 in either a standalone read-based approach or a combined approach in the case of population-level data. Phasing allows for the partitioning of sequencing reads from the original, unmasked alignments into haplotype subsets, which might then be evaluated separately for allele-specific methylation using a conventional approach such as MethylDackel (<https://github.com/dpryan79/MethylDackel>).

Furthermore, the method might be integrated further with inferences from stable epigenetic marks such as 5mC in CG context. There is currently no available software which links read-based variants and 5mC positions with population-scale data, which might lead to a marked improvement in estimation of allele-specific methylation, particularly when all inferences are drawn in the first place from the same exact sequencing libraries. Evaluation of such a method might involve the use of model test species, such as *A. thaliana*, where high-quality haplotype references already exist for direct comparison.

8 Conclusion

This thesis discusses current advances in the computational analysis of epigenetic marks from NGS data, in its application to plant ecology; notably including the relevance of a high-quality reference genome, as demonstrated with a hybrid assembly approach, and introducing several new pipelines for the downstream analysis of bisulfite sequencing data. Present understanding of the mechanisms and functional consequences of plant epigenetics is derived largely from model species such as *Arabidopsis thaliana*, which has proven historically to be an excellent model to study the genetic basis of plant evolution. The breadth of comparison between model and non-model organisms in regards to their epigenetic processes however remains to be seen, particularly in the context of evolutionary ecology. (Richards et al. 2017) write about the need to better integrate the fields of molecular genetics and evolutionary ecology, by adding more ecological context and ecological questions to model species research (e.g. Latzel et al. 2013; Hagmann et al. 2015), and by adopting higher resolution tools in non-model species research (e.g. Platt et al. 2015; Xie et al. 2015; Gugger et al. 2016; van Gulp et al. 2016; Trucchi et al. 2016). Under the purview of the EpiDiverse project (<https://epidiverse.eu>) this thesis attempts to address the latter topic, first in a specific sense by providing new resources for the non-model plant species *Thlaspi arvense*, and second in a broader sense by developing new pipelines and analysis tools to facilitate high-resolution analysis in a wide range of non-model species, including *T. arvense*.

The annual weed, field pennycress (*T. arvense*), is a Brassicacea closely-related to *A. thaliana*, but with a larger genome showing markedly different patterns in terms of extensive repeat content and its inherent level of DNA methylation. The genomic resources presented herein provide a basis for example in the analysis of natural populations, wherein more than 200 individuals have already been collected by EpiDiverse across Europe to investigate the extent and relevance of its epigenetic variation (Galanti et al. 2022). This large-scale analysis also makes use of the pipelines and computational tools provided herein. Furthermore, with the placement of pennycress as an emerging crop species, researchers can utilise these resources to aid efforts in crop breeding and domestication.

With a high-quality reference genome in place, focus turned towards the application of bisulfite sequencing to study genome-wide patterns of DNA methylation and the main bioinformatic tasks

8 Conclusion

associated herewith. A comprehensive benchmark of nine, bisulfite-aware, short-read alignment tools revealed performance differences in precision-recall and downstream consequences pertaining to methylation level deviations in difficult-to-map regions. Performant tools were selected for inclusion in a new computational pipeline for quality control, alignment and methylation level quantification in a highly-parallel manner. Also benchmarked were a number of tools for variant calling in bisulfite sequencing data, including a new method developed and implemented as part of this thesis. This new “double-masking” procedure was shown to outperform all existing methods in terms of precision-recall, in the case of publicly available benchmark datasets for both human and a model plant species. In addition, the software deviates from existing tools in that it is instead a pre-processing procedure which makes it possible to leverage state-of-the-art variant calling tools developed for conventional sequencing data, thereby removing the dependency on specialised variant callers altogether. Researchers familiar with popular tools such as GATK and FreeBayes are able to continue using them now with bisulfite sequencing data. The capability to accurately discriminate between genetic and epigenetic variation using the same sequencing library is of great practical value to ecologists studying non-model species, particularly at a population-level, as the cost of sequencing both conventional and bisulfite libraries is often prohibitive. The method was also implemented as part of a collaboration to build an updated and revised epiGBS2 pipeline, a separate software for investigating methylation patterns in reduced representation bisulfite sequencing data, where consensus reference sequences are constructed *de novo* in cases where an appropriate reference genome is not available. The sensitivity in RRBS data is lower due to inherent strand bias present in such sequencing libraries, but the precision remains comparably high as in WGBS data. This approach represents another useful tool for plant ecologists studying epigenetics.

Moreover, several other downstream analyses for bisulfite sequencing data were also considered in their application to non-model plant species on a population-scale. Current methods for differential methylation analysis were investigated and compared for suitability under different experimental conditions. Most prevalent throughout the field are those developed for case/control studies, for example when investigating the transgenerational effect of stress-response or exposure to various environmental conditions. Identification of differentially methylated regions remains a powerful tool when inferring transcriptional and regulatory consequences of DNA methylation patterns, but there is much variation in how they should be consistently defined under specific circumstances. Given that the priority is most often to generate candidate regions for further study, the sensitivity of methylation was considered to be advantageous when it comes to non-model plant data, particularly

in light of its robust segmentation approach and non-parametric statistical test; unlike some other methods it avoids drawing assumptions from methylation patterns in other species. Alternatively, a new method MethylScore uses an unsupervised, HMM approach to identify DMRs without prior knowledge of sample structure. This presents an opportunity for population-scale data, where interacting environmental conditions and/or complex phenotypes often obscure associated methylation patterns amongst the background variation, and opens the door for a “reverse epigenetics” approach to help resolve hidden relationships at a local level. Both methylene and MethylScore were implemented in easy-to-use pipelines for use by plant ecologists.

In addition, a pipeline for performing epigenome-wide association studies with non-model plant data was implemented, making use of the existing R package Gene, Environment and Methylation (GEM) and expanding its features in regards to both improved missing data imputation and mitigation of the heavy multiple testing burden common in such analyses. Given a quantitative trait or environmental condition measured within the confines of the experimental design, EWAS can identify significant associations with epigenetic markers among population data thus allowing for the elucidation of specific molecular mechanisms. The pipeline is interoperable with the output of all other relevant pipelines developed during the course of this thesis, including i) sample-specific methylation calls which are the main basis of the analysis, ii) genetic variants which allow for testing the interaction between different kinds of associations, and iii) input of DMRs as markers which can provide a stronger signal for comparison and necessitate fewer multiple comparisons. The WGBS, DMR, SNP and EWAS pipelines together comprise the core of the “EpiDiverse Toolkit” (Figure 35), the main outcome of this thesis alongside the genome assembly for pennycress.

Finally, further improvements in the implementation of the “double-masking” procedure for variant calling were explored, in the context of deriving haplotypes for allele-specific methylation (ASM). As a read-based approach both provides context for local phasing while simultaneously carrying methylation information for calculating ASM, an integrated method may provide greater precision than existing methods which rely on data from independent sequencing libraries. By incorporating linkage disequilibrium analysis derived from population-level data sets, and/or stable epigenetic marks, construction of even longer haplotypes may be possible. An integrated approach which includes all these aspects has not yet been attempted in the literature, and could offer a powerful alternative to existing methods which is appropriate for both model and non-model species.

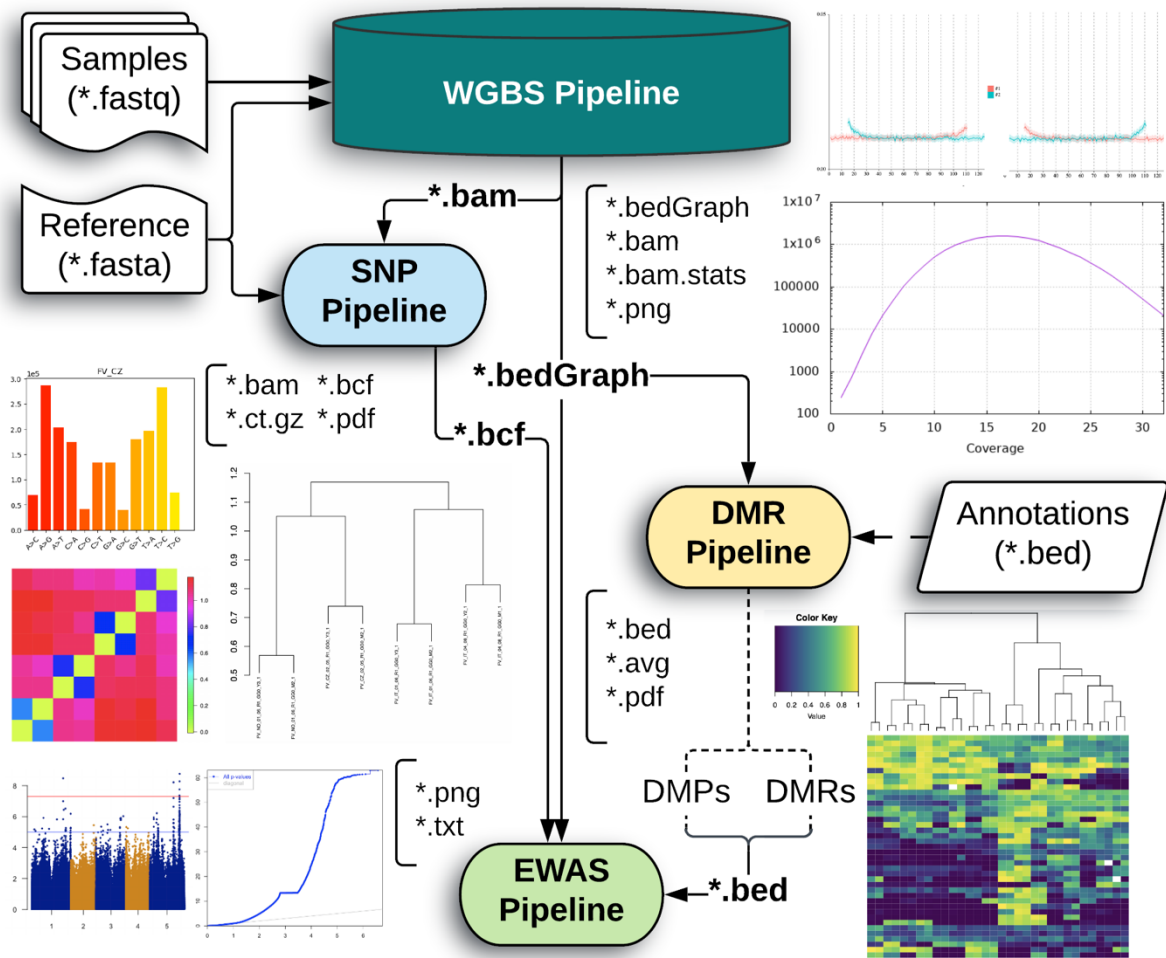
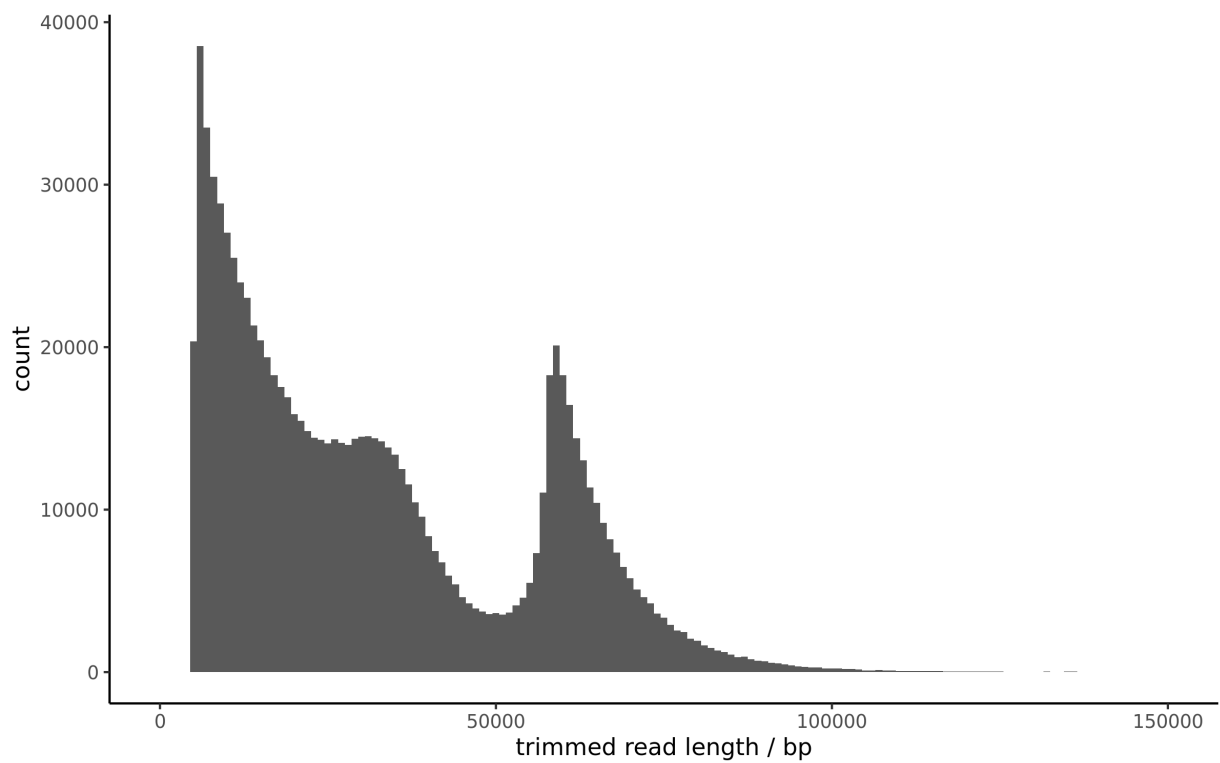


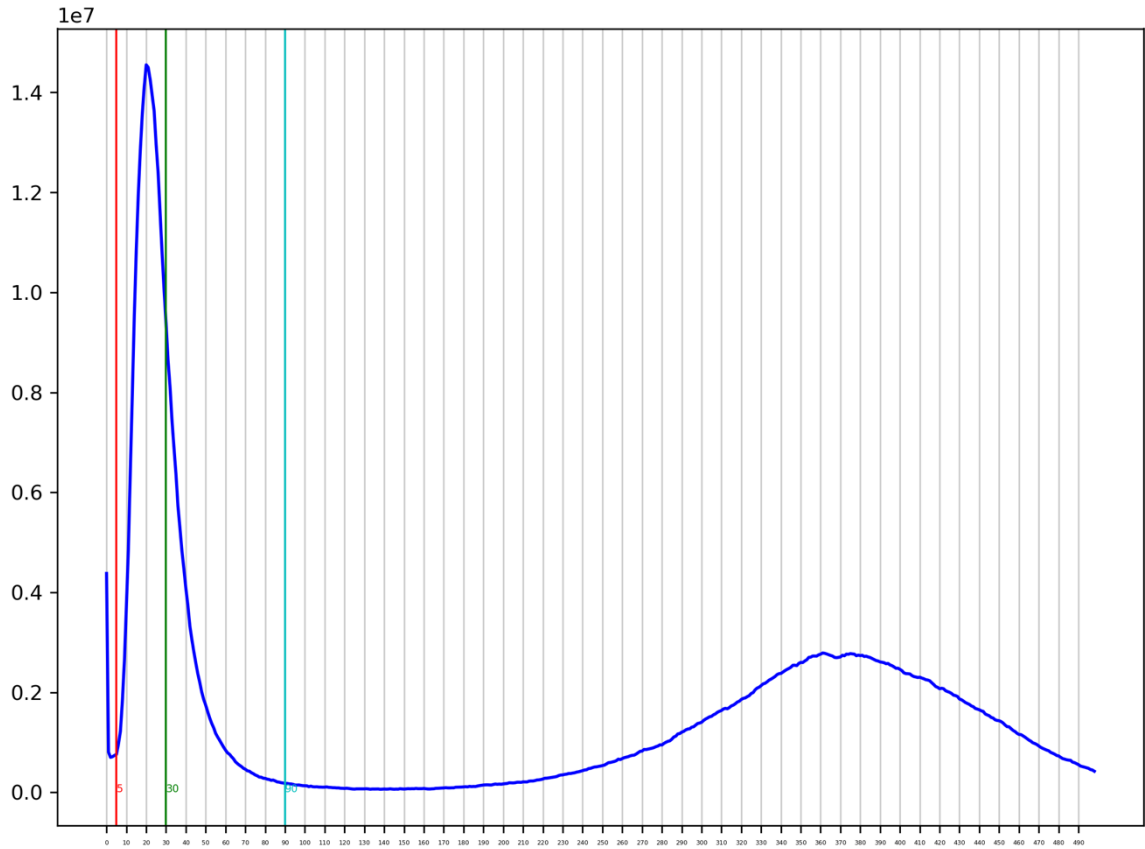
Figure 35. An overview of the EpiDiverse Toolkit. The WGBS data forms the foundation of the analysis, and each downstream pipeline is built to work either in cooperation with one another or, optionally, with independently-generated input data. All pipelines output runtime metadata, tracing and further visualisation in addition to what is shown here. The full output is described for each pipeline in the documentation hosted on Github at <https://github.com/EpiDiverse>

Appendices

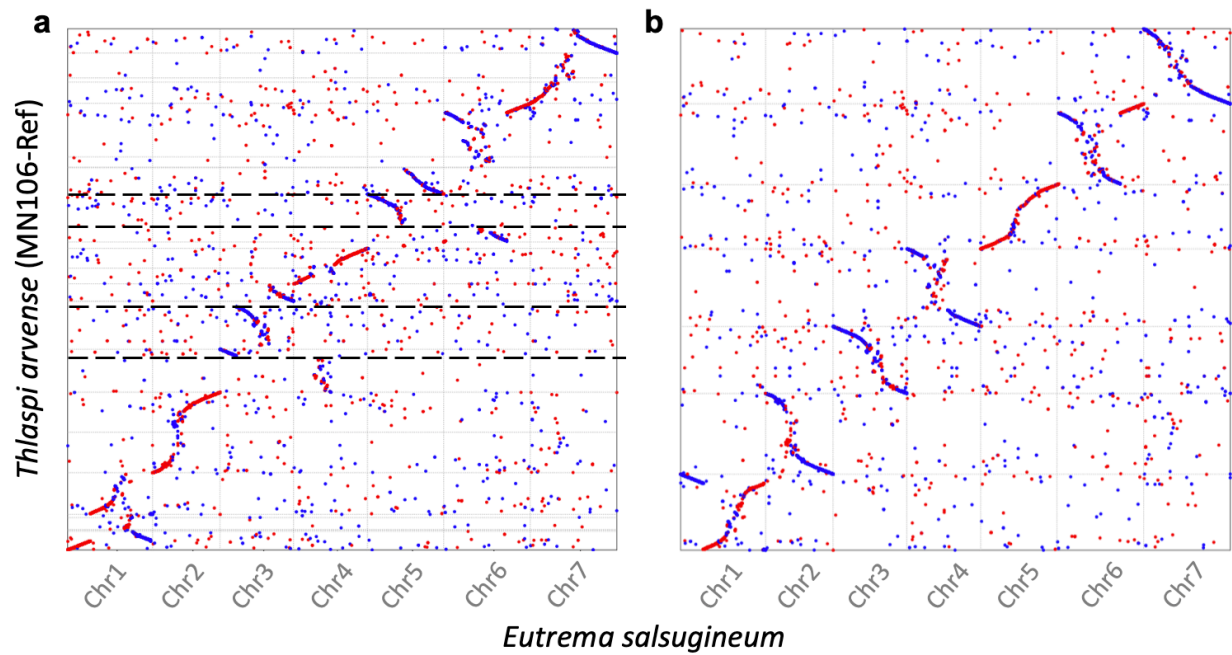
A. Supplement: Building a Suitable Reference Genome



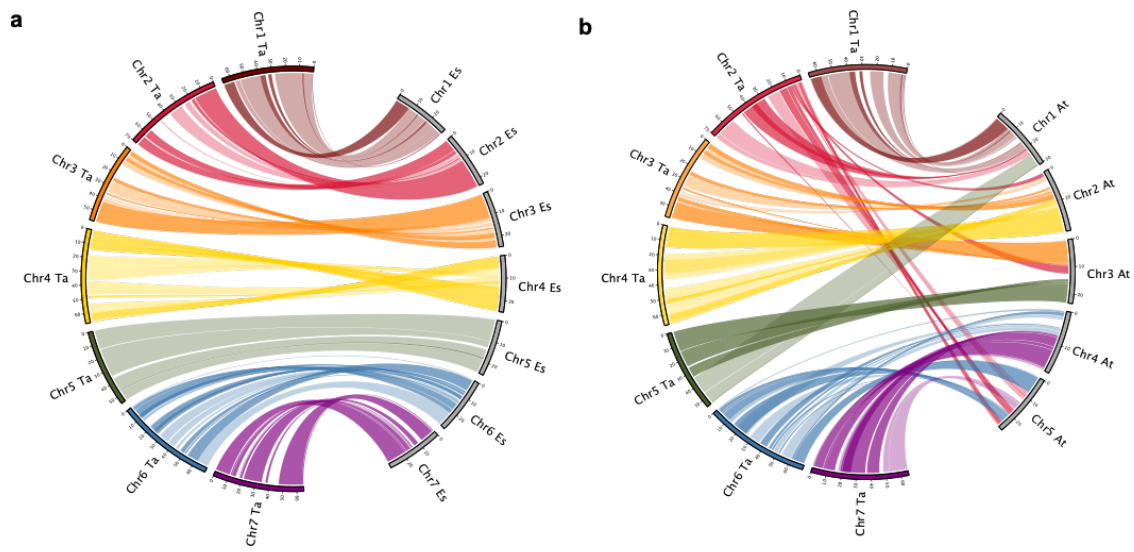
Supplementary Figure A.1 Read length distribution of trimmed PacBio Sequel II CLR reads taken forward for assembly with Canu v1.9.



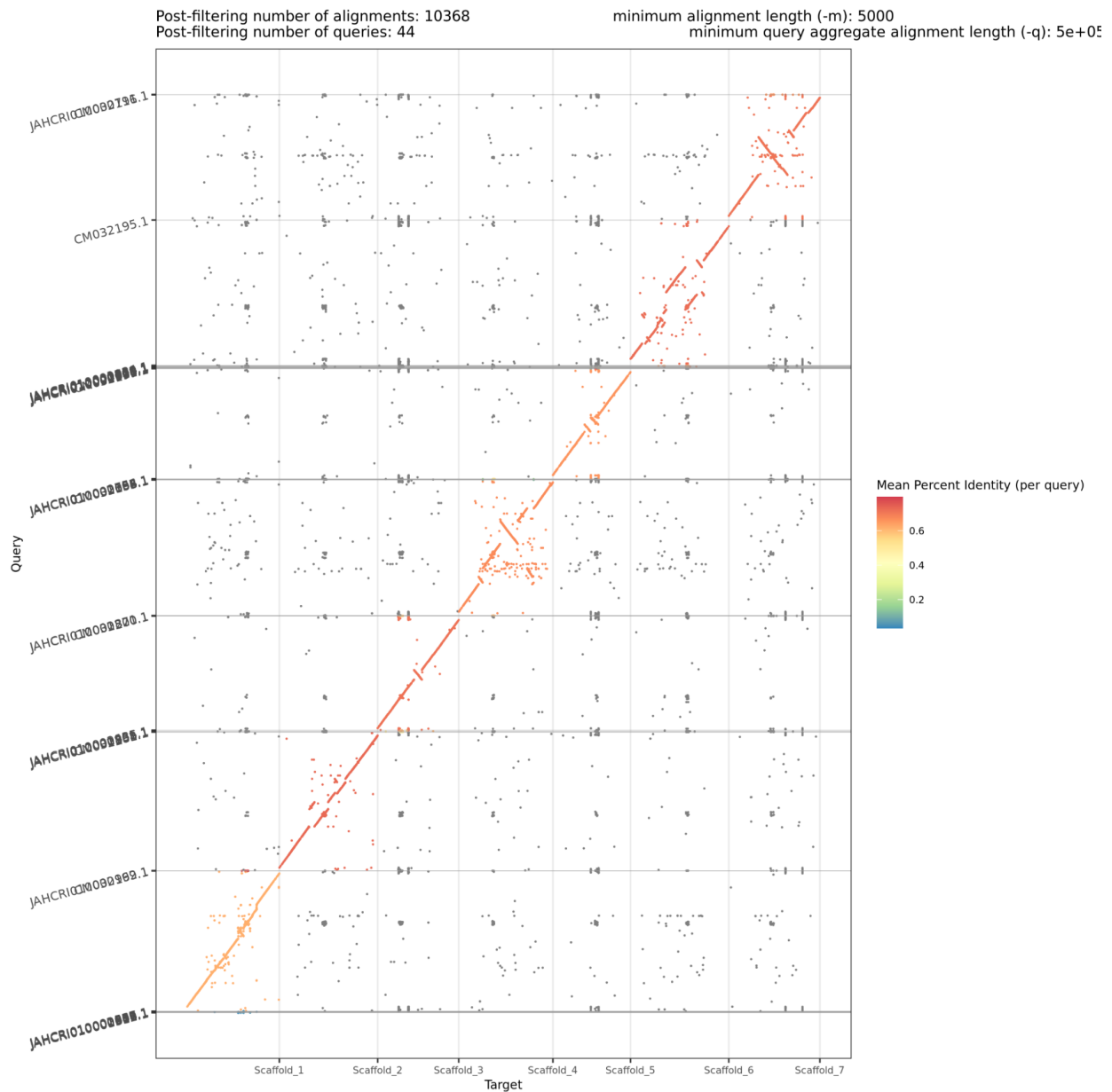
Supplementary Figure A.2 Distribution of PacBio Sequel II CLR read mapping depth frequency over assembled contigs, with bimodal peaks due to contig regions with lower depth than the average indicating that they are duplicated.



Supplementary Figure A.3 Sequence dot plots showing the largest seven scaffolds of the closely-related species *E. salsugineum* and their equivalent in *T. arvense* var. MN106-Ref (T_arvense_v2), comparing the difference both **a)** before and **b)** after re-scaffolding. Horizontal dashed lines in **(a)** denote breakpoints which were manually introduced to the genome based on evaluation of genetic maps, synteny maps, Hi-C data, and comparison with YUN_Tarv1.0.



Supplementary Figure A.4 Synteny analysis between the largest seven scaffolds of *T. arvense* var. MN106-Ref (Ta) and **a**) their equivalent in the closely-related species *E. salsugineum* (Es), and **b**) *A. thaliana* (At). The Ribbons show the syntenic relationships between the two genomes. Dark ribbons indicate syntenic blocks in inverse orientation.



Supplementary Figure A.5 Synteny between *T_arvense_v2* (x-axis) and *YUN_Tarv_1.0* (y-axis). Significantly more un-scaffolded contigs from *YUN_Tarv_1.0* map to *T_arvense_v2* than vice versa, with a total of 44 query sequences from the Chinese accession retained after post-filtering for minimum alignment length 5,000 and aggregate alignment length of 500,000. Most un-scaffolded contigs map to pericentromeric and centromeric regions, which are visible here in addition to the notable mis-scaffolding of centromeric repeats at the telomeric ends of the chromosome-representing scaffolds in *YUN_Tarv_1.0*.

Supplementary Table A.1 Estimation of the genome size of *T. arvense* using flow cytometry with *Arabidopsis thaliana*, tomato (*Solanum lycopersicum*), and maize (*Zea mays*) as references.

Sample	DNA Content (pg)	Predicted Genome Size using the reference (Mbp)
Field pennycress (MN106-Ref)	1.09	NA
<i>Arabidopsis thaliana</i>	0.32	459
Tomato	2.05	505
Maize	5.45	540

Supplementary Table A.2 Full descriptive statistics for intermediate versions of the assembly starting with correction, trimming and initial assembly of PacBio reads (Canu), further polishing and scaffolding using optical maps and contact maps (Bionano + HiC), and the final version following manual curation and re-scaffolding with the help of genetic linkage and synteny maps (ALLMAPS).

	Canu	Bionano + HiC	ALLMAPS
# contigs / scaffolds	4,704	976	964
≥ 1 Kbp	4,704	976	964
≥ 5 Kbp	4,704	976	964
≥ 10 Kbp	4,582	965	953
≥ 25 Kbp	4,027	902	890
≥ 50 Kbp	2,750	619	607
Total length	797 Mbp	526 Mbp	526 Mbp
≥ 1 Kbp	797 Mbp	526 Mbp	526 Mbp
≥ 5 Kbp	797 Mbp	526 Mbp	526 Mbp
≥ 10 Kbp	796 Mbp	525 Mbp	525 Mbp
≥ 25 Kbp	786 Mbp	524 Mbp	524 Mbp
≥ 50 Kbp	737 Mbp	514 Mbp	514 Mbp
Largest scaffold	64.4 Mbp	64.4 Mbp	70.0 Mbp
N50	15.0 Mbp	34.7 Mbp	64.9 Mbp
NG50	34.7 Mbp	34.7 Mbp	64.9 Mbp
N75	0.11 Mbp	15.0 Mbp	61.0 Mbp
NG75	15.0 Mbp	15.0 Mbp	55.2 Mbp
L50	13	7	4
LG50	7	7	4
L75	884	13	6
LG75	13	13	7
GC (%)	37.08	38.39	38.39
# N's per 100 Kbp	0.25	0.29	0.51

Supplementary Table A.3 BUSCO statistics on **a)** initial assembly, immediately after CANU, and **b)** final assembly. Both are derived from orthologs to the Eudicotyledons odb10 database.

a) C:98.4%[S:74.8%,D:23.6%],F:0.6%,M:1.0%,n:2121	
2086	Complete BUSCOs (C)
1586	Complete and single-copy BUSCOs (S)
500	Complete and duplicated BUSCOs (D)
12	Fragmented BUSCOs (F)
23	Missing BUSCOs (M)
2121	Total BUSCO groups searched
b) C:98.7%[S:92.1%,D:6.6%],F:0.5%,M:0.8%,n:2121	
2094	Complete BUSCOs (C)
1954	Complete and single-copy BUSCOs (S)
140	Complete and duplicated BUSCOs (D)
11	Fragmented BUSCOs (F)
16	Missing BUSCOs (M)
2121	Total BUSCO groups searched

Supplementary Table A.4 Merqury k-mer (k=21) analysis of Illumina HiSeq reads sequenced from the accession in YUN_Tarv_1.0, showing greater QV scores in T_arvense_v2 for the equivalent top 7 scaffolds based on k-mers found uniquely in each assembly and those shared with the read set.

Scaffold	T_arvense_v2				YUN_Tarv_1.0			
	length	uniq. k-mer	QV	error	length	uniq. k-mer	QV	error
1	65,519,694	1,288,483	30.24	0.0009	73,273,813	3,665,265	26.13	0.0024
2	70,024,556	1,619,266	29.53	0.0011	72,401,710	2,823,932	27.23	0.0019
3	63,812,002	969,655	31.37	0.0007	59,618,324	2,615,059	26.71	0.0021
4	60,964,055	1,833,722	28.38	0.0015	70,763,934	2,808,229	27.15	0.0019
5	55,234,666	1,400,039	29.13	0.0012	57,454,197	3,020,585	25.90	0.0026
6	69,981,056	2,370,907	27.85	0.0016	75,772,189	4,364,339	25.50	0.0028
7	64,850,309	1,826,069	28.67	0.0014	65,366,657	2,922,800	26.62	0.0022

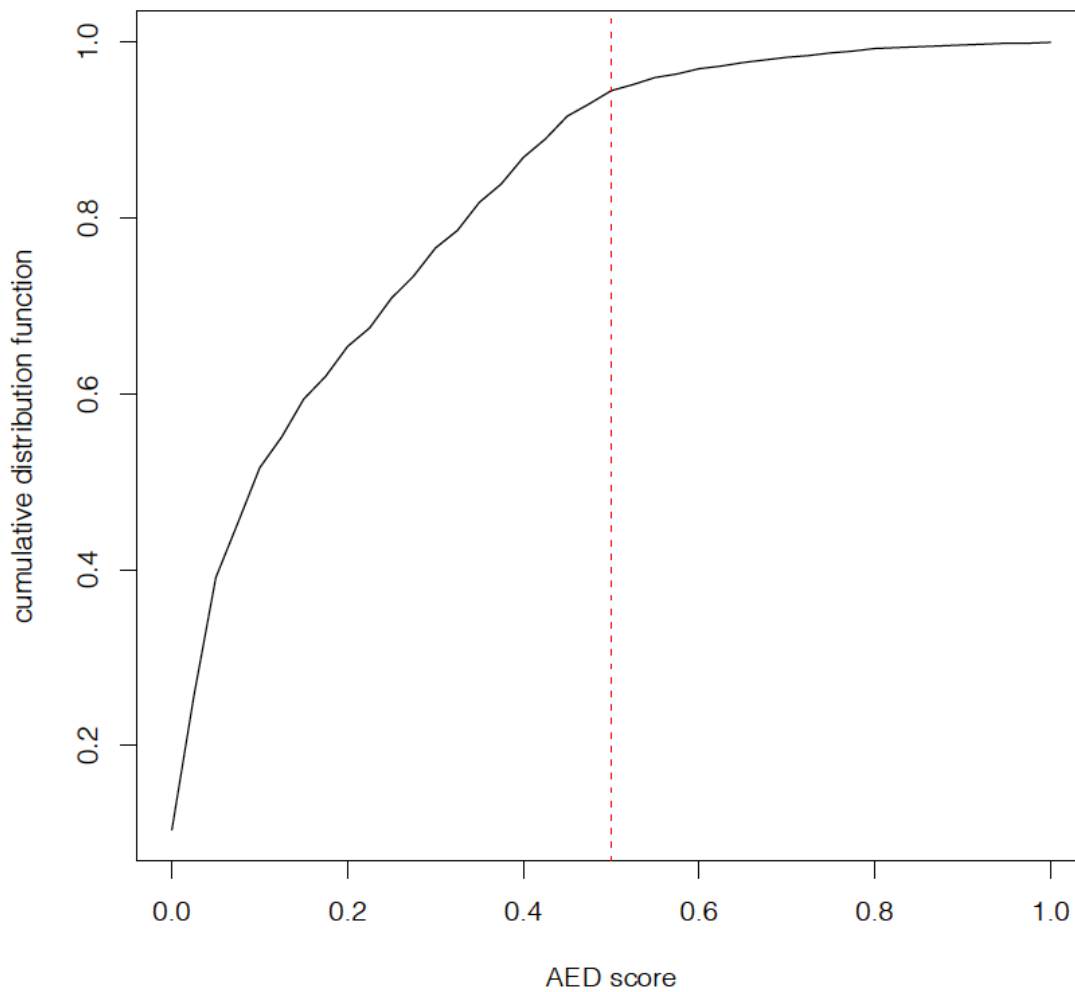
Supplementary Table A.5 Merqury k-mer (k=21) analysis of Illumina HiSeq reads (PCR-free) sequenced from the accession MN106-Ref, showing greater QV scores in T_arvense_v2 for the equivalent top 7 scaffolds based on k-mers found uniquely in each assembly and those shared with the read set.

Scaffold	T_arvense_v2				YUN_Tarv_1.0			
	length	uniq. k-mer	QV	error	length	uniq. k-mer	QV	error
1	65,519,694	98,643	41.44	7.2E-05	73,273,813	4,505,381	25.20	0.0030
2	70,024,556	78,174	42.74	5.3E-05	72,401,710	4,121,308	25.55	0.0028
3	63,812,002	109,566	40.87	8.2E-05	59,618,324	3,133,918	25.90	0.0026
4	60,964,055	69,169	42.67	5.4E-05	70,763,934	4,393,318	25.16	0.0030
5	55,234,666	41,707	44.44	3.6E-05	57,454,197	4,022,088	24.62	0.0035
6	69,981,056	71,047	43.15	4.8E-05	75,772,189	5,931,889	24.12	0.0039
7	64,850,309	74,777	42.60	5.5E-05	65,366,657	4,436,822	24.76	0.0033

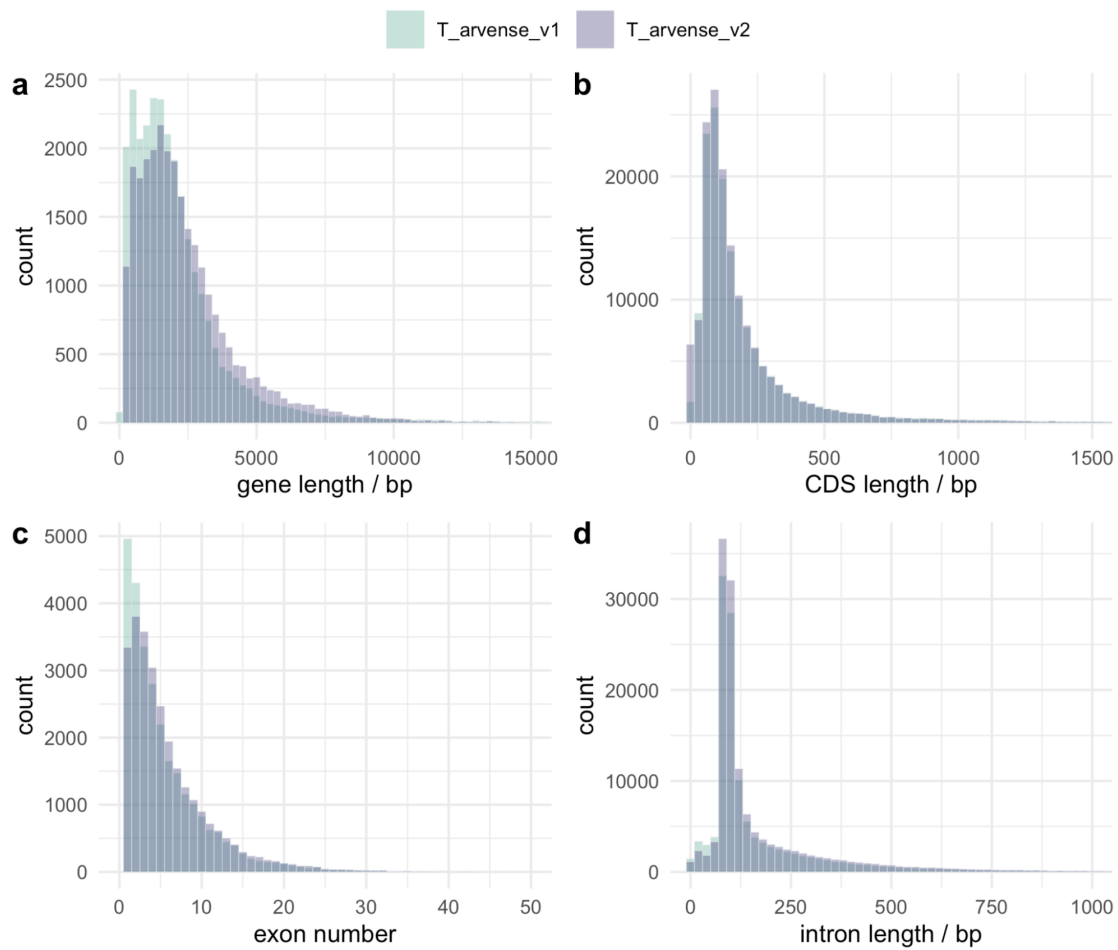
Supplementary Table A.6 Merqury k-mer (k=21) analysis of each total assembly showing relative completeness of k-mers present in each read set from Illumina HiSeq.

Assembly	Read set	k-mers (asm)	k-mers (reads)	% completeness
YUN_Tarv_1.0	MN106-Ref	211,016,166	228,654,390	92.2861
T_arvense_v2	MN106-Ref	227,350,203	228,654,390	99.4296
both	MN106-Ref	227,710,775	228,654,390	99.5873
YUN_Tarv_1.0	SRR14757813	223,058,386	229,823,560	97.0564
T_arvense_v2	SRR14757813	215,415,116	229,823,560	93.7306
both	SRR14757813	227,717,433	229,823,560	99.0836

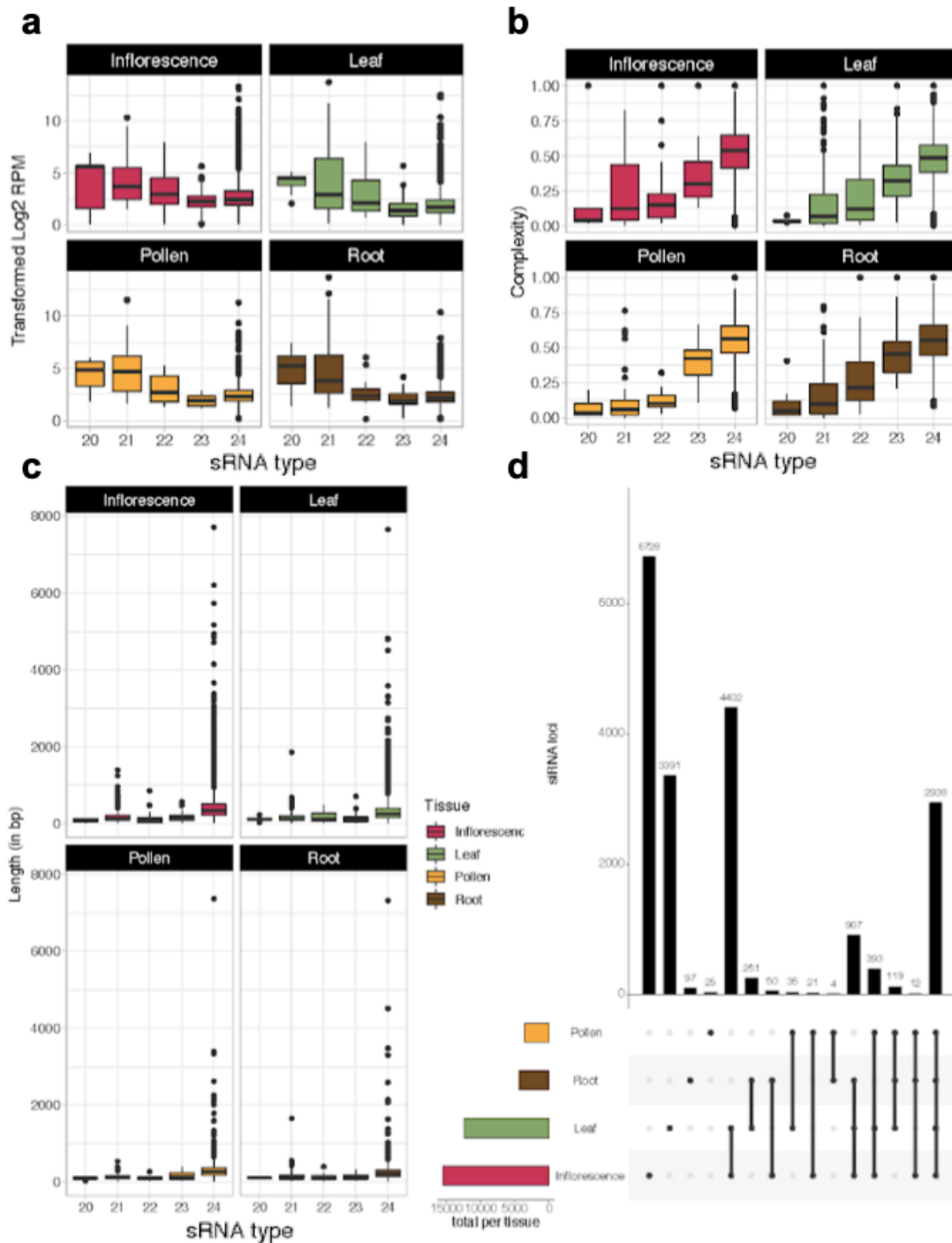
B. Supplement: Feature Annotation for Epigenomics



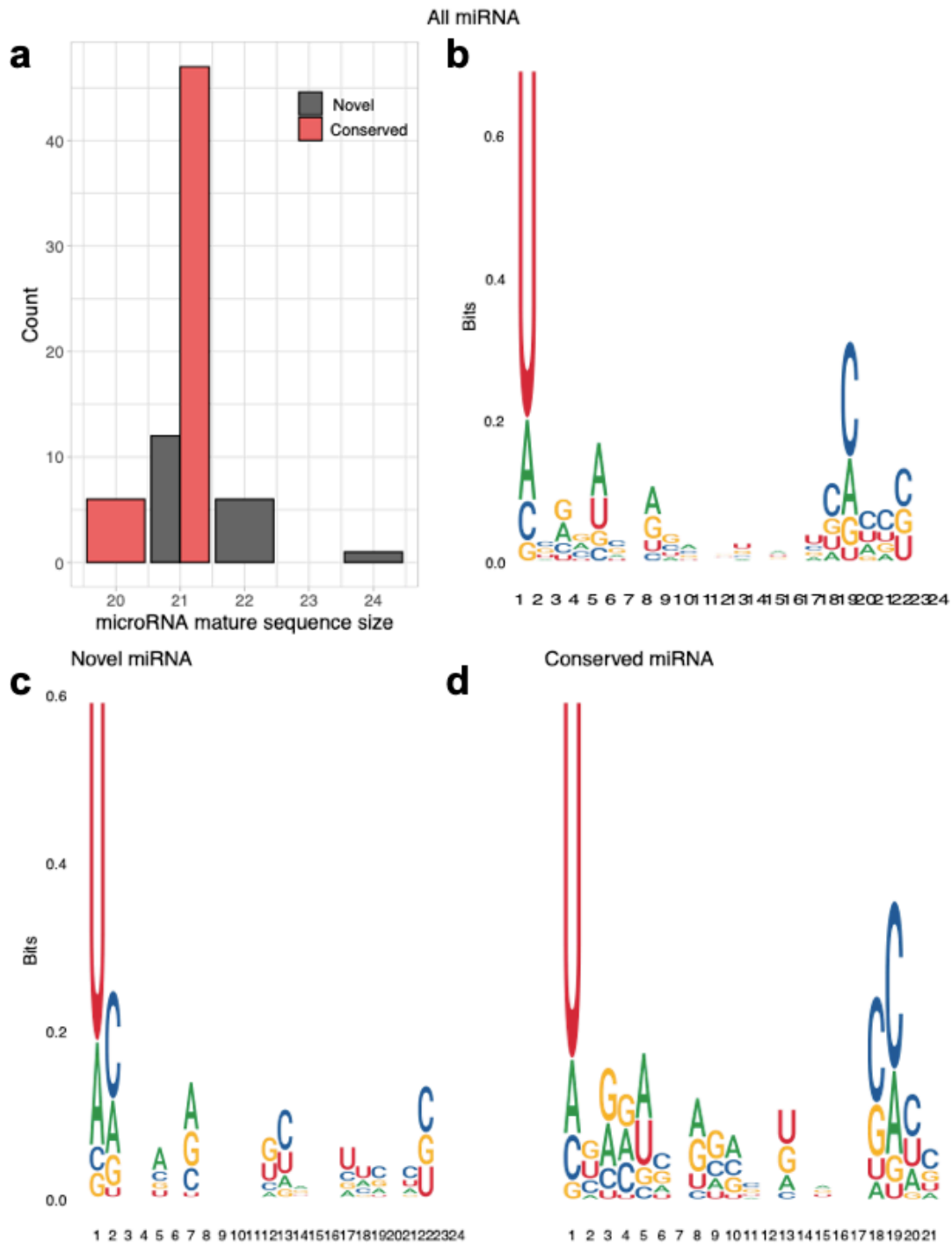
Supplementary Figure B.1 The cumulative distribution of annotation edit distance (AED) scores from the final set of protein-coding loci, denoting that ~95% of annotated genes are supported with a score ≤ 0.5 overall.



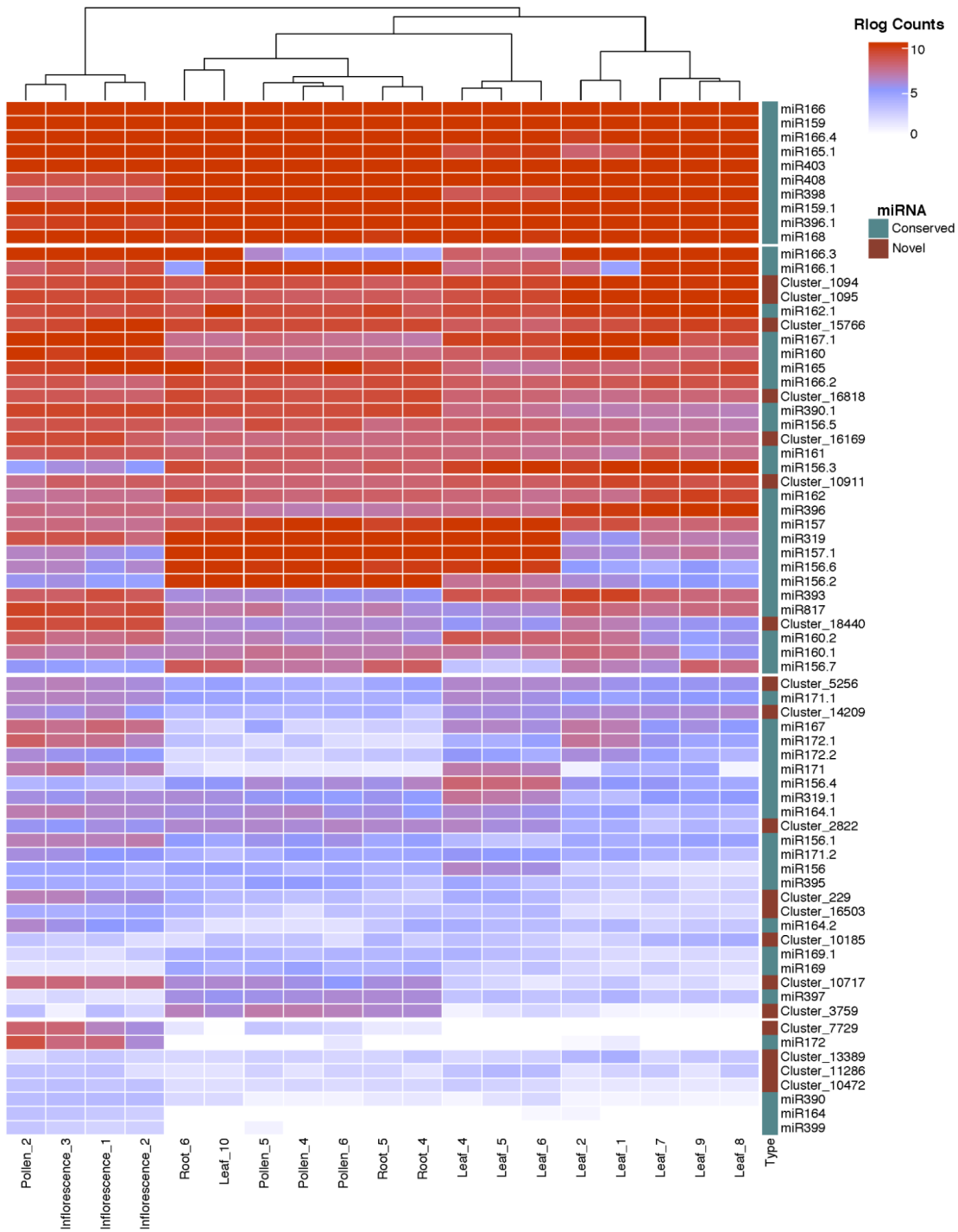
Supplementary Figure B.2 An overview of annotated genomic feature distributions in comparison to T_arvense_v1 for **a)** gene lengths, **b)** CDS lengths, **c)** per gene exon number, and **d)** intron lengths.



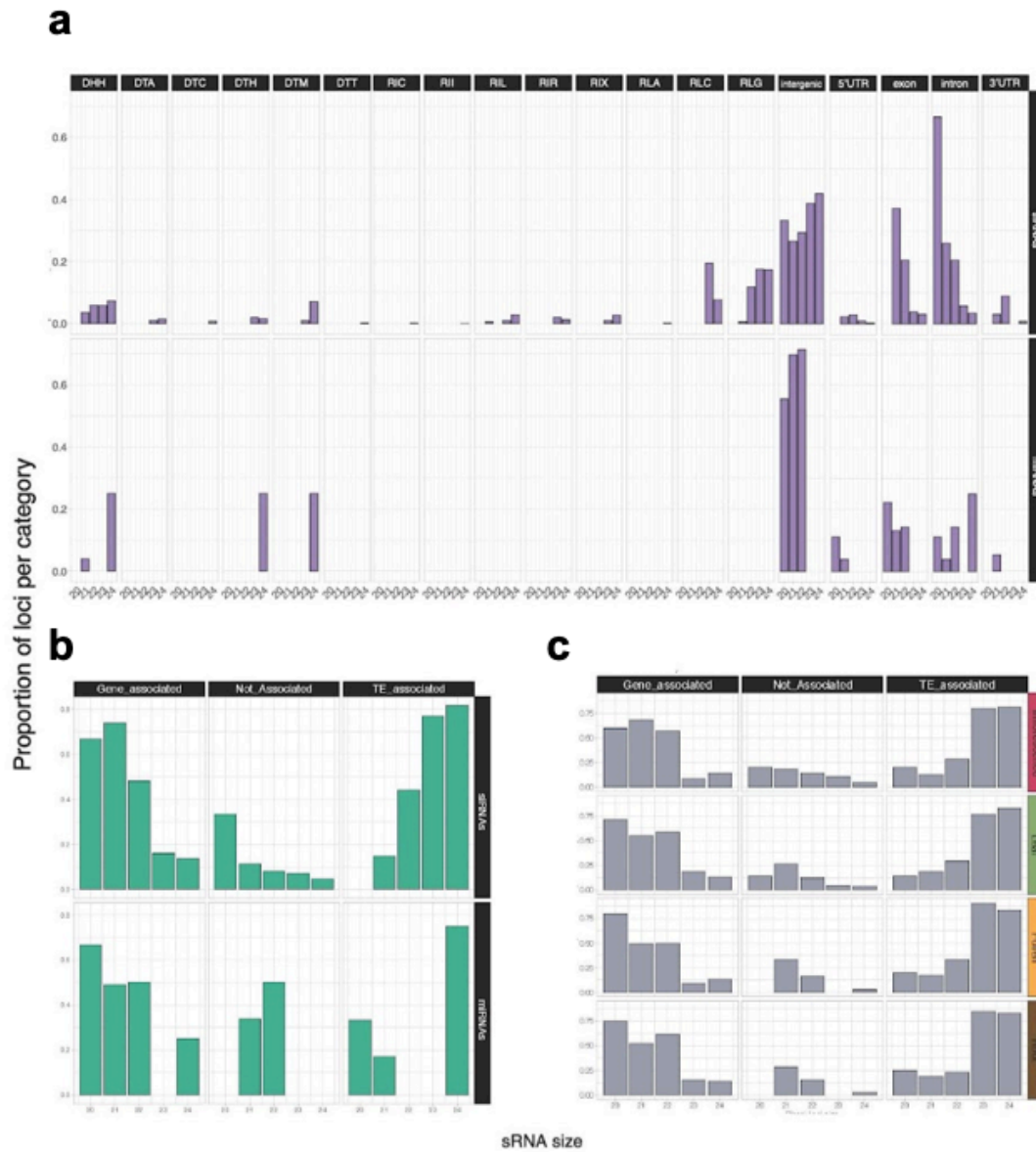
Supplementary Figure B.3 Small RNA (sRNA) annotation in the *T_arvense_v2* genome assembly. **a)** sRNA loci per tissue of origin (RPM = reads per million). **b)** sRNA complexity, measured as “number of distinct alignments / total number of alignments”. **c)** sRNA locus size distribution. **d)** Co-occurrence of sRNAs between tissues. Coloured horizontal bars show the total number of loci per tissue.



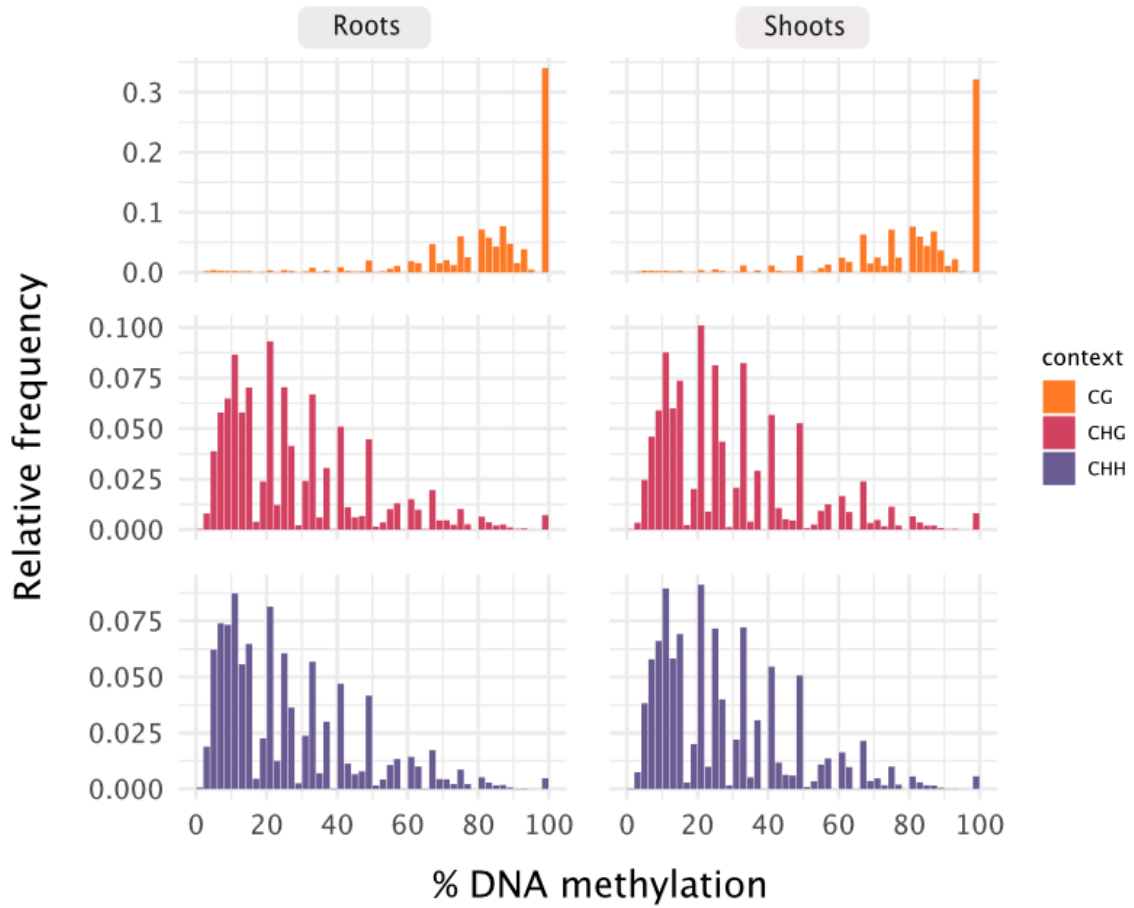
Supplementary Figure B.4 Predicted miRNAs in the *T_arvensis_v2* genome assembly. **a)** Size of mature miRNAs identified in this study, split into novel and conserved miRNA species. **b-d)** Sequence conservation in miRNAs, measured in bits for each position of the mature microRNA (Schneider and Stephens 1990). Sequences were aligned from the 5'-end. Different panels show sequence conservation of all miRNAs (**b**), only novel miRNAs (**c**) and only conserved miRNAs (**d**).



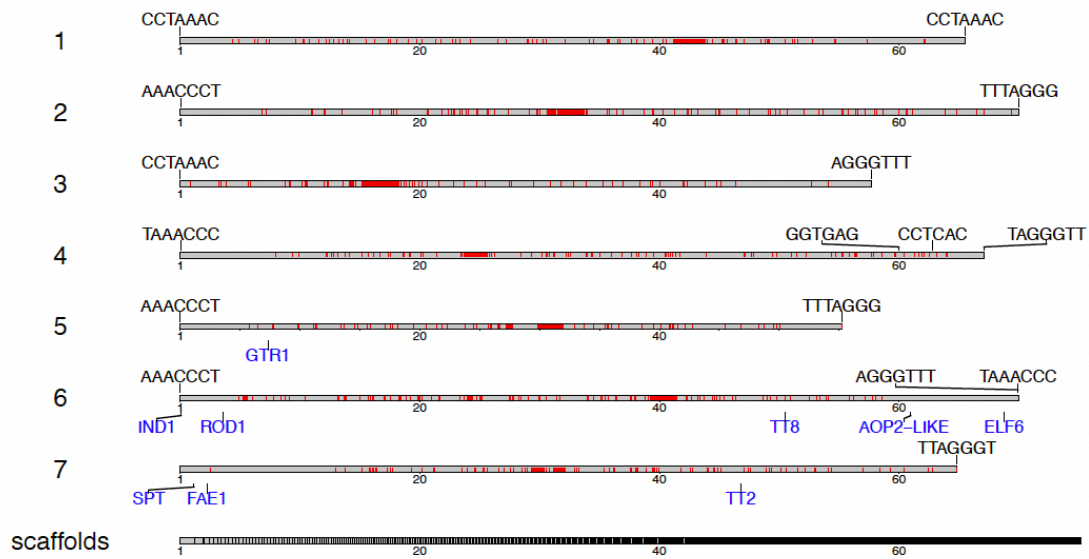
Supplementary Figure B.5 Relative expression level of novel and conserved miRNA families between tissue types. Sample counts are reported as transformed count data on the \log_2 scale which has been normalised with respect to library size.



Supplementary Figure B.6 sRNA types and their association with different genomic features. **a)** Occupancy of all annotated siRNAs and miRNAs in either genes, TE superfamilies, or intergenic regions. **b,c)** Association of sRNA loci with either TEs or genes within 1.5 Kb distance, for all sRNAs (**b**) and for only phased loci (**c**).



Supplementary Figure B.7 Methylation rate frequency distribution by sequence context in shoot and root tissues.



Supplementary Figure B.8 Karyotype plot of the seven largest scaffolds representing chromosomes in *T. arvensis* MN106-Ref (*T_arvensis_v2*), alongside a concatenation of all minor scaffolds. Transposable element LTR annotations (red) are highlighted as an approximate localisation of each centromere, alongside loci containing putative telomeric repeat motifs (black labels) and genes of interest (blue labels) in the *de novo* domestication of pennycress. Scaling is given in Mbp.

Supplementary Table B.1 Alignment statistics of mRNA-seq reads prior to merging by tissue type.

sample ID	tissue	replicate	# reads	# alignments	% mapping rate
92479	Cauline leaf	1	26,807,084	22,791,298	85.02
92480	Cauline leaf	2	30,493,022	28,502,974	93.47
92468	Cauline leaf	3	34,703,622	32,482,275	93.6
92460	Green seed	1	34,409,210	30,592,247	88.91
92451	Green seed	2	4,350,126	731,192	16.81
92476	Green seed	3	15,881,224	10,421,117	65.62
92454	Inflorescence	1	37,241,578	35,012,072	94.01
92455	Inflorescence	2	44,709,756	42,480,499	95.01
92456	Inflorescence	3	40,602,958	38,486,094	94.79
92471	Mature seed	1	8,563,360	3,050,475	35.62
92459	Mature seed	2	8,788,058	2,824,810	32.14
92475	Old green silique	2	6,472,836	1,831,577	28.3
92467	Old green silique	3	6,097,296	955,425	15.67
92470	Old green silique	4	8,450,100	3,160,559	37.4
92482	Open flowers	1	27,784,846	26,166,091	94.17
92452	Open flowers	2	48,974,428	40,861,167	83.43
92453	Open flowers	3	33,257,622	31,907,812	95.94
92473	Root 1 week old	1	32,394,826	29,934,857	92.41
92474	Root 1 week old	2	29,742,158	27,530,520	92.56
92458	Root 1 week old	3	33,174,194	31,436,791	94.76
92469	Rosette leaf	1	28,171,012	26,329,294	93.46
92478	Rosette leaf	2	27,902,144	26,281,165	94.19
92481	Rosette leaf	3	31,754,984	30,136,304	94.9
92461	Seed pod	1	39,817,146	36,813,142	92.46
92462	Seed pod	2	37,222,832	34,593,888	92.94
92477	Seed pod	3	27,609,204	24,962,517	90.41
92466	Shoot 1 week old	1	41,735,500	38,909,856	93.23
92472	Shoot 1 week old	2	32,111,696	30,143,229	93.87
92457	Shoot 1 week old	4	35,692,836	34,067,138	95.45
92463	Young green silique	1	37,909,974	34,540,606	91.11
92464	Young green silique	2	36,102,364	32,464,393	89.92
92465	Young green silique	3	35,177,290	30,480,559	86.65

C. Supplement: From Read Alignment to DNA Methylation Analysis

Supplementary Table C.1 Test system specifications.

Operating System	CentOS Linux 7 (Core)
Architecture	x86_64
CPU Model	Intel(R) Xeon(R) Gold 6130
Clock Speed	2.10 GHz
Available CPUs	64
Available RAM	256 Gb
File System(s)	File ext4

D. Supplement: Inferring Genomic Information

Supplementary Table D.1 The command line executed for each variant calling software. Commands were executed as jobs on a HPC cluster. Examples are given for *Arabidopsis thaliana*, and in all cases the variable **\$1** denotes the input alignment (*.bam) file.

Software	Command Line
BISCUIT	<pre>biscuit pileup -m 1 -b 1 -q 1 \ -o \$(basename \$1 .bam).vcf genome/arabidopsis.fa \$1</pre>
Bis-SNP	<pre>java -Xmx10g -jar bin/BisSNP-1.0.1.jar \ -R genome/arabidopsis.fa \ -T BisulfiteGenotyper \ -I \$1 \ -vfn1 \$(basename \$1 .bam).vcf \ -out_modes EMIT_VARIANTS_ONLY \ -mmq 1 -mbq 1</pre>
BS-SNPer	<pre>perl bin/BS-Snper/BS-Snper.pl \ --fa genome/arabidopsis.fa \ --input \$1 \ --output \$(basename \$1 .bam).snps \ --methcg \$(basename \$1 .bam).methcg \ --methchg \$(basename \$1 .bam).methchg \ --methchh \$(basename \$1 .bam).methchh \ --minhetfreq 0.1 \ --minhomfreq 0.85 \ --minquali 1 \ --mincover 1 \ --maxcover 1000 \ --minread2 2 \ --errorate 0.02 \ --mapvalue 1 > \$(basename \$1 .bam).vcf</pre>
MethylExtract	<pre>perl bin/MethylExtract.pl \ seq=genome/arabidopsis.fa \</pre>

```

inDir=$1 \
minQ=1 \
varFraction=0.05 \
maxPval=0.01 \
outDir=methylextract \
flagW=0 \
flagC=16 # or flagW=99,147 flagC=83,163 for PE

```

FreeBayes

```

freebayes -f genome/arabidopsis.fa $1 \
--no-partial-observations \
--report-genotype-likelihood-max \
--genotype-qualities \
--min-repeat-entropy 1 \
--min-coverage 1 \
--min-base-quality 1 > $(basename $1 .bam).vcf

```

**GATK3.8
UnifiedGenotyper**

```

gatk3 -T UnifiedGenotyper \
-R genome/arabidopsis.fa \
-I $1 \
-o $(basename $1 .bam).vcf.gz \
--min_base_quality_score 1

```

**Platypus
(standard mode)**

```

platypus callVariants \
--refFile=genome/arabidopsis.fa \
--bamFiles=$1 \
--output=$(basename $1 .bam).vcf \
--minMapQual=1 \
--minBaseQual=1

```

**Platypus
(assembly mode)**

```

platypus callVariants \
--refFile=genome/arabidopsis.fa \
--bamFiles=$1 \
--output=$(basename $1 .bam).vcf \
--assemble=1 \
--assembleBadReads=1 \
--minMapQual=1 \
--minBaseQual=0

```


Supplementary Table D.2 Summary of QC metrics as described by GATK at the time of publication, from <https://gatk.broadinstitute.org/hc/en-us/articles/360035890471> (accessed 8th January 2020). Filtering thresholds are according to best-practices.

Metric	Threshold	Summary
Quality (QUAL)	< 30	Variant confidence. The Phred-scaled probability that there is some kind of nucleotide variation at a given site.
QualByDepth (QD)	< 2	The variant confidence (from the QUAL field) divided by the unfiltered depth of non-hom-ref samples. Intended to normalize the variant quality in order to avoid inflation caused when there is deep coverage.
FisherStrand (FS)	> 60	The Phred-scaled probability that there is strand bias at the site, indicating whether the alternate allele was seen more or less often on the forward or reverse strand than the reference allele.
StrandOddsRatio (SOR)	> 3	Another way to estimate strand bias using a test similar to the symmetric odds ratio test. Created because FS tends to penalise variants that occur at the ends of exons.
RMSMappingQuality (MQ)	< 40	The root mean square mapping quality over all the reads at a given site.
MappingQualityRankSumTest (MQRankSum)	< -12.5	The u-based z-approximation from the Rank Sum Test for mapping qualities. It compares the mapping qualities of the reads supporting the reference allele and the alternate allele.
ReadPosRankSumTest (ReadPosRankSum)	< -8	The u-based z-approximation from the Rank Sum Test for site position within reads. It compares whether the positions of the reference and alternate alleles are different within the reads.

Bibliography

1000 Genomes Project Consortium, Adam Auton, Lisa D. Brooks, Richard M. Durbin, Erik P. Garrison, Hyun Min Kang, Jan O. Korbel, et al. 2015. “A Global Reference for Human Genetic Variation.” *Nature* 526 (7571): 68–74.

1001 Genomes Consortium. 2016. “1,135 Genomes Reveal the Global Pattern of Polymorphism in *Arabidopsis Thaliana*.” *Cell* 166 (2): 481–91.

Aird, Daniel, Michael G. Ross, Wei-Sheng Chen, Maxwell Danielsson, Timothy Fennell, Carsten Russ, David B. Jaffe, Chad Nusbaum, and Andreas Gnirke. 2011. “Analyzing and Minimizing PCR Amplification Bias in Illumina Sequencing Libraries.” *Genome Biology* 12 (2): R18.

Akalin, Altuna, Matthias Kormaksson, Sheng Li, Francine E. Garrett-Bakelman, Maria E. Figueroa, Ari Melnick, and Christopher E. Mason. 2012. “methylKit: A Comprehensive R Package for the Analysis of Genome-Wide DNA Methylation Profiles.” *Genome Biology* 13 (10): R87.

Alkan, Can, Saba Sajjadian, and Evan E. Eichler. 2011. “Limitations of next-Generation Genome Sequence Assembly.” *Nature Methods* 8 (1): 61–65.

Alonso, Conchita, Daniela Ramos-Cruz, and Claude Becker. 2019. “The Role of Plant Epigenetics in Biotic Interactions.” *The New Phytologist* 221 (2): 731–37.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. “Basic Local Alignment Search Tool.” *Journal of Molecular Biology* 215 (3): 403–10.

Amselem, Joëlle, Guillaume Cornut, Nathalie Choisne, Michael Alaux, Françoise Alfama-Depauw, Véronique Jamilloux, Florian Maumus, et al. 2019. “RepetDB: A Unified Resource for Transposable Element References.” *Mobile DNA* 10 (January): 6.

Andersen, Tonni Grube, and Barbara Ann Halkier. 2014. “Upon Bolting the GTR1 and GTR2 Transporters Mediate Transport of Glucosinolates to the Inflorescence rather than Roots.” *Plant Signaling & Behavior* 9 (1): e27740.

- Atwell, Susanna, Yu S. Huang, Bjarni J. Vilhjálmsson, Glenda Willems, Matthew Horton, Yan Li, Dazhe Meng, et al. 2010. "Genome-Wide Association Study of 107 Phenotypes in Arabidopsis Thaliana Inbred Lines." *Nature* 465 (7298): 627–31.
- Aufsatz, Werner, M. Florian Mette, Johannes van der Winden, Antonius J. M. Matzke, and Marjori Matzke. 2002. "RNA-Directed DNA Methylation in Arabidopsis." *Proceedings of the National Academy of Sciences of the United States of America* 99 Suppl 4 (December): 16499–506.
- Axtell, Michael J. 2013a. "Classification and Comparison of Small RNAs from Plants." *Annual Review of Plant Biology* 64 (January): 137–59.
- . 2013b. "ShortStack: Comprehensive Annotation and Quantification of Small RNA Genes." *RNA* 19 (6): 740–51.
- Bao, Weidong, Kenji K. Kojima, and Oleksiy Kohany. 2015. "Repbase Update, a Database of Repetitive Elements in Eukaryotic Genomes." *Mobile DNA* 6 (June): 11.
- Barturen, Guillermo, Antonio Rueda, José L. Oliver, and Michael Hackenberg. 2013. "MethylExtract: High-Quality Methylation Maps and SNV Calling from Whole Genome Bisulfite Sequencing Data." *F1000Research* 2 (October): 217.
- Bastow, Ruth, Joshua S. Mylne, Clare Lister, Zachary Lippman, Robert A. Martienssen, and Caroline Dean. 2004. "Vernalization Requires Epigenetic Silencing of FLC by Histone Methylation." *Nature* 427 (6970): 164–67.
- Becker, Claude, Jörg Hagmann, Jonas Müller, Daniel Koenig, Oliver Stegle, Karsten Borgwardt, and Detlef Weigel. 2011. "Spontaneous Epigenetic Variation in the Arabidopsis Thaliana Methyloome." *Nature* 480 (7376): 245–49.
- Beilstein, Mark A., Nathalie S. Nagalingum, Mark D. Clements, Steven R. Manchester, and Sarah Mathews. 2010. "Dated Molecular Phylogenies Indicate a Miocene Origin for Arabidopsis Thaliana." *Proceedings of the National Academy of Sciences of the United States of America* 107 (43): 18724–28.
- Benjamini, Yuval, and Terence P. Speed. 2012. "Summarizing and Correcting the GC Content Bias in High-Throughput Sequencing." *Nucleic Acids Research* 40 (10): e72.
- Benson, G. 1999. "Tandem Repeats Finder: A Program to Analyze DNA Sequences." *Nucleic Acids Research* 27 (2): 573–80.

Bibliography

- Bentley, David R., Shankar Balasubramanian, Harold P. Swerdlow, Geoffrey P. Smith, John Milton, Clive G. Brown, Kevin P. Hall, et al. 2008. "Accurate Whole Human Genome Sequencing Using Reversible Terminator Chemistry." *Nature* 456 (7218): 53–59.
- Beric, Aleksandra, Makenzie E. Mabry, Alex E. Harkess, Julia Brose, M. Eric Schranz, Gavin C. Conant, Patrick P. Edger, Blake C. Meyers, and J. Chris Pires. 2021. "Comparative Phylogenetics of Repetitive Elements in a Diverse Order of Flowering Plants (Brassicales)." *G3*, May. <https://doi.org/10.1093/g3journal/jkab140>.
- Bewick, Adam J., Lexiang Ji, Chad E. Niederhuth, Eva-Maria Willing, Brigitte T. Hofmeister, Xiuling Shi, Li Wang, et al. 2016. "On the Origin and Evolutionary Consequences of Gene Body DNA Methylation." *Proceedings of the National Academy of Sciences of the United States of America* 113 (32): 9111–16.
- Bewick, Adam J., and Robert J. Schmitz. 2017. "Gene Body DNA Methylation in Plants." *Current Opinion in Plant Biology* 36 (April): 103–10.
- Boateng, A. A., C. A. Mullen, and N. M. Goldberg. 2010. "Producing Stable Pyrolysis Liquids from the Oil-Seed Presscakes of Mustard Family Plants: Pennycress (*Thlaspi Arvense* L.) and Camelina (*Camelina Sativa*)." *Energy & Fuels: An American Chemical Society Journal* 24 (12): 6624–32.
- Bormann Chung, Christina A., Victoria L. Boyd, Kevin J. McKernan, Yutao Fu, Cinna Monighetti, Heather E. Peckham, and Melissa Barker. 2010. "Whole Methylome Analysis by Ultra-Deep Sequencing Using Two-Base Encoding." *PloS One* 5 (2): e9320.
- Bossdorf, Oliver, Davide Arcuri, Christina L. Richards, and Massimo Pigliucci. 2010. "Experimental Alteration of DNA Methylation Affects the Phenotypic Plasticity of Ecologically Relevant Traits in *Arabidopsis Thaliana*." *Evolutionary Ecology* 24 (3): 541–53.
- Bossdorf, Oliver, Christina L. Richards, and Massimo Pigliucci. 2008. "Epigenetics for Ecologists." *Ecology Letters* 11 (2): 106–15.
- Boutet, Emmanuel, Damien Lieberherr, Michael Tognolli, Michel Schneider, and Amos Bairoch. 2007. "UniProtKB/Swiss-Prot." *Methods in Molecular Biology* 406: 89–112.
- Boyko, Alex, Todd Blevins, Youli Yao, Andrey Golubov, Andriy Bilichak, Yaroslav Ilnytskyy, Jens Hollunder, Frederick Meins Jr, and Igor Kovalchuk. 2010. "Transgenerational Adaptation of *Arabidopsis* to Stress Requires DNA Methylation and the Function of Dicer-like Proteins." *PloS One* 5 (3): e9514.

- Bradnam, Keith R., Joseph N. Fass, Anton Alexandrov, Paul Baranay, Michael Bechner, Inanç Birol, Sébastien Boisvert, et al. 2013. “Assemblathon 2: Evaluating de Novo Methods of Genome Assembly in Three Vertebrate Species.” *GigaScience* 2 (1): 10.
- Browning, Sharon R., and Brian L. Browning. 2011. “Haplotype Phasing: Existing Methods and New Developments.” *Nature Reviews. Genetics* 12 (10): 703–14.
- Bucher, Etienne, Jon Reinders, and Marie Mirouze. 2012. “Epigenetic Control of Transposon Transcription and Mobility in Arabidopsis.” *Current Opinion in Plant Biology* 15 (5): 503–10.
- Buckler, Edward S., James B. Holland, Peter J. Bradbury, Charlotte B. Acharya, Patrick J. Brown, Chris Browne, Elhan Ersoz, et al. 2009. “The Genetic Architecture of Maize Flowering Time.” *Science* 325 (5941): 714–18.
- Bush, William S., and Jason H. Moore. 2012. “Chapter 11: Genome-Wide Association Studies.” *PLoS Computational Biology* 8 (12): e1002822.
- Campbell, Michael S., Carson Holt, Barry Moore, and Mark Yandell. 2014. “Genome Annotation and Curation Using MAKER and MAKER-P.” *Current Protocols in Bioinformatics / Editorial Board, Andreas D. Baxevanis ... [et Al.]* 48 (1): 4.11.1–39.
- Campbell, Michael S., Meiyee Law, Carson Holt, Joshua C. Stein, Gaurav D. Moghe, David E. Hufnagel, Jikai Lei, et al. 2014. “MAKER-P: A Tool Kit for the Rapid Creation, Management, and Quality Control of Plant Genome Annotations.” *Plant Physiology* 164 (2): 513–24.
- 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.” *Current Biology: CB* 13 (24): 2212–17.
- Cao, Xiaofeng, and Steven E. Jacobsen. 2002. “Role of the Arabidopsis DRM Methyltransferases in de Novo DNA Methylation and Gene Silencing.” *Current Biology: CB* 12 (13): 1138–44.
- Catoni, Marco, Thomas Jonesman, Elisa Cerruti, and Jerzy Paszkowski. 2019. “Mobilization of Pack-CACTA Transposons in Arabidopsis Suggests the Mechanism of Gene Shuffling.” *Nucleic Acids Research* 47 (3): 1311–20.
- Cervera, M. T., L. Ruiz-García, and J. M. Martínez-Zapater. 2002. “Analysis of DNA Methylation in Arabidopsis Thaliana Based on Methylation-Sensitive AFLP Markers.” *Molecular Genetics and Genomics: MGG* 268 (4): 543–52.

Bibliography

- Chakraborty, Mahul, James G. Baldwin-Brown, Anthony D. Long, and J. J. Emerson. 2016. "Contiguous and Accurate de Novo Assembly of Metazoan Genomes with Modest Long Read Coverage." *Nucleic Acids Research* 44 (19): e147.
- Chatterjee, Aniruddha, Peter A. Stockwell, Euan J. Rodger, and Ian M. Morison. 2012. "Comparison of Alignment Software for Genome-Wide Bisulphite Sequence Data." *Nucleic Acids Research* 40 (10): e79.
- Cheng, Jingfei, Qingfeng Niu, Bo Zhang, Kunsong Chen, Ruihua Yang, Jian-Kang Zhu, Yijing Zhang, and Zhaobo Lang. 2018. "Downregulation of RdDM during Strawberry Fruit Ripening." *Genome Biology* 19 (1): 212.
- Chen, Pao-Yang, Shawn J. Cokus, and Matteo Pellegrini. 2010. "BS Seeker: Precise Mapping for Bisulfite Sequencing." *BMC Bioinformatics* 11 (April): 203.
- Cherf, Gerald M., Kate R. Lieberman, Hytham Rashid, Christopher E. Lam, Kevin Karplus, and Mark Akeson. 2012. "Automated Forward and Reverse Ratcheting of DNA in a Nanopore at 5-Å Precision." *Nature Biotechnology* 30 (4): 344–48.
- Chong, Suyinn, and Emma Whitelaw. 2004. "Epigenetic Germline Inheritance." *Current Opinion in Genetics & Development* 14 (6): 692–96.
- Chopra, Ratan, Nicole Folstad, Joseph Lyons, Tim Ulmasov, Cynthia Gallaher, Liam Sullivan, Abby McGovern, et al. 2019. "The Adaptable Use of Brassica NIRS Calibration Equations to Identify Pennycress Variants to Facilitate the Rapid Domestication of a New Winter Oilseed Crop." *Industrial Crops and Products* 128 (February): 55–61.
- Chopra, Ratan, Nicole Folstad, and M. David Marks. 2020. "Combined Genotype and Fatty-Acid Analysis of Single Small Field Pennycress (*Thlaspi Arvense*) Seeds Increases the Throughput for Functional Genomics and Mutant Line Selection." *Industrial Crops and Products* 156 (November): 112823.
- Chopra, Ratan, Evan B. Johnson, Erin Daniels, Michaela McGinn, Kevin M. Dorn, Maliheh Esfahanian, Nicole Folstad, et al. 2018. "Translational Genomics Using Arabidopsis as a Model Enables the Characterization of Pennycress Genes through Forward and Reverse Genetics." *The Plant Journal: For Cell and Molecular Biology* 96 (6): 1093–1105.

- Chopra, Ratan, Evan B. Johnson, Ryan Emenecker, Edgar B. Cahoon, Joe Lyons, Daniel J. Kliebenstein, Erin Daniels, et al. 2020. "Identification and Stacking of Crucial Traits Required for the Domestication of Pennycress." *Nature Food* 1 (1): 84–91.
- Claver, Ana, Raquel Rey, M. Victoria López, Rafael Picorel, and Miguel Alfonso. 2017. "Identification of Target Genes and Processes Involved in Erucic Acid Accumulation during Seed Development in the Biodiesel Feedstock Pennycress (*Thlaspi Arvense* L.)." *Journal of Plant Physiology* 208 (January): 7–16.
- Cleary, John G., Ross Braithwaite, Kurt Gaastra, Brian S. Hilbush, Stuart Inglis, Sean A. Irvine, Alan Jackson, et al. 2015. "Comparing Variant Call Files for Performance Benchmarking of Next-Generation Sequencing Variant Calling Pipelines." *bioRxiv*. <https://doi.org/10.1101/023754>.
- 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." *Nature* 452 (7184): 215–19.
- Cortijo, Sandra, René Wardenaar, Maria Colomé-Tatché, Arthur Gilly, Mathilde Etcheverry, Karine Labadie, Erwann Cailieux, et al. 2014. "Mapping the Epigenetic Basis of Complex Traits." *Science* 343 (6175): 1145–48.
- Cubas, P., C. Vincent, and E. Coen. 1999. "An Epigenetic Mutation Responsible for Natural Variation in Floral Symmetry." *Nature* 401 (6749): 157–61.
- Cubins, Julija A., M. Scott Wells, Katherine Frels, Matthew A. Ott, Frank Forcella, Gregg A. Johnson, Maninder K. Walia, Roger L. Becker, and Russ W. Gesch. 2019. "Management of Pennycress as a Winter Annual Cash Cover Crop. A Review." *Agronomy for Sustainable Development* 39 (5): 46.
- Danecek, Petr, James K. Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O. Pollard, Andrew Whitwham, et al. 2021. "Twelve Years of SAMtools and BCFtools." *GigaScience* 10 (2). <https://doi.org/10.1093/gigascience/giab008>.
- Deaton, Aimée M., and Adrian Bird. 2011. "CpG Islands and the Regulation of Transcription." *Genes & Development* 25 (10): 1010–22.
- Delaneau, Olivier, Cédric Coulonges, and Jean-François Zagury. 2008. "Shape-IT: New Rapid and Accurate Algorithm for Haplotype Inference." *BMC Bioinformatics* 9 (December): 540.

Bibliography

- Delaneau, Olivier, Bryan Howie, Anthony J. Cox, Jean-François Zagury, and Jonathan Marchini. 2013. "Haplotype Estimation Using Sequencing Reads." *American Journal of Human Genetics* 93 (4): 687–96.
- Delaneau, Zagury, Robinson, Marchini, and Dermitzakis. 2019. "Accurate, Scalable and Integrative Haplotype Estimation." *Nature Communications*. <https://www.nature.com/articles/s41467-019-13225-y>.
- De La Torre, Amanda R., Inanc Birol, Jean Bousquet, Pär K. Ingvarsson, Stefan Jansson, Steven J. M. Jones, Christopher I. Keeling, et al. 2014. "Insights into Conifer Giga-Genomes." *Plant Physiology* 166 (4): 1724–32.
- De Lucia, Filomena, Pedro Crevillen, Alexandra M. E. Jones, Thomas Greb, and Caroline Dean. 2008. "A PHD-Polycomb Repressive Complex 2 Triggers the Epigenetic Silencing of FLC during Vernalization." *Proceedings of the National Academy of Sciences of the United States of America* 105 (44): 16831–36.
- Denton, James F., Jose Lugo-Martinez, Abraham E. Tucker, Daniel R. Schrider, Wesley C. Warren, and Matthew W. Hahn. 2014. "Extensive Error in the Number of Genes Inferred from Draft Genome Assemblies." *PLoS Computational Biology* 10 (12): e1003998.
- Di Tommaso, Paolo, Maria Chatzou, Evan W. Floden, Pablo Prieto Barja, Emilio Palumbo, and Cedric Notredame. 2017. "Nextflow Enables Reproducible Computational Workflows." *Nature Biotechnology* 35 (4): 316–19.
- Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. 2013. "STAR: Ultrafast Universal RNA-Seq Aligner." *Bioinformatics* 29 (1): 15–21.
- Dorn, Kevin M., Johnathon D. Fankhauser, Donald L. Wyse, and M. David Marks. 2013. "De Novo Assembly of the Pennycress (*Thlaspi Arvense*) Transcriptome Provides Tools for the Development of a Winter Cover Crop and Biodiesel Feedstock." *The Plant Journal: For Cell and Molecular Biology* 75 (6): 1028–38.
- . 2015. "A Draft Genome of Field Pennycress (*Thlaspi Arvense*) Provides Tools for the Domestication of a New Winter Biofuel Crop." *DNA Research: An International Journal for Rapid Publication of Reports on Genes and Genomes* 22 (2): 121–31.

- Dubin, Manu J., Pei Zhang, Dazhe Meng, Marie-Stanislas Remigereau, Edward J. Osborne, Francesco Paolo Casale, Philipp Drewe, et al. 2015. "DNA Methylation in Arabidopsis Has a Genetic Basis and Shows Evidence of Local Adaptation." *eLife* 4 (May): e05255.
- Du, Huilong, and Chengzhi Liang. 2019. "Assembly of Chromosome-Scale Contigs by Efficiently Resolving Repetitive Sequences with Long Reads." *Nature Communications* 10 (1): 5360.
- Du, Jiamu, Xuehua Zhong, Yana V. Bernatavichute, Hume Stroud, Suhua Feng, Elena Caro, Ajay A. Vashisht, et al. 2012. "Dual Binding of Chromomethylase Domains to H3K9me2-Containing Nucleosomes Directs DNA Methylation in Plants." *Cell* 151 (1): 167–80.
- Du, Xiongming, Gai Huang, Shoupu He, Zhaoen Yang, Gaofei Sun, Xiongfeng Ma, Nan Li, et al. 2018. "Resequencing of 243 Diploid Cotton Accessions Based on an Updated A Genome Identifies the Genetic Basis of Key Agronomic Traits." *Nature Genetics* 50 (6): 796–802.
- Eberle, Carrie A., Matthew D. Thom, Kristine T. Nemecek, Frank Forcella, Jonathan G. Lundgren, Russell W. Gesch, Walter E. Riedell, et al. 2015. "Using Pennycress, Camelina, and Canola Cash Cover Crops to Provision Pollinators." *Industrial Crops and Products* 75 (November): 20–25.
- Edger, Patrick P., Robert VanBuren, Marivi Colle, Thomas J. Poorten, Ching Man Wai, Chad E. Niederhuth, Elizabeth I. Alger, et al. 2018. "Single-Molecule Sequencing and Optical Mapping Yields an Improved Genome of Woodland Strawberry (*Fragaria Vesca*) with Chromosome-Scale Contiguity." *GigaScience* 7 (2): 1–7.
- Eichten, Steve R., Ruth A. Swanson-Wagner, James C. Schnable, Amanda J. Waters, Peter J. Hermanson, Sanzhen Liu, Cheng-Ting Yeh, et al. 2011. "Heritable Epigenetic Variation among Maize Inbreds." *PLoS Genetics* 7 (11): e1002372.
- Eid, John, Adrian Fehr, Jeremy Gray, Khai Luong, John Lyle, Geoff Otto, Paul Peluso, et al. 2009. "Real-Time DNA Sequencing from Single Polymerase Molecules." *Science* 323 (5910): 133–38.
- Emms, David M., and Steven Kelly. 2017. "STRIDE: Species Tree Root Inference from Gene Duplication Events." *Molecular Biology and Evolution* 34 (12): 3267–78.
- Emms, D. M., and S. Kelly. 2018. "STAG: Species Tree Inference from All Genes." *bioRxiv*. <https://doi.org/10.1101/267914>.
- Erdmann, Robert M., Amanda L. Souza, Clary B. Clish, and Mary Gehring. 2014. "5-Hydroxymethylcytosine Is Not Present in Appreciable Quantities in Arabidopsis DNA." *G3* 5 (1): 1–8.

Bibliography

- Esfahanian, Maliheh, Tara J. Nazareus, Meghan M. Freund, Gary McIntosh, Winthrop B. Phippen, Mary E. Phippen, Timothy P. Durrett, Edgar B. Cahoon, and John C. Sedbrook. 2021. “Generating Pennycress (*Thlaspi Arvense*) Seed Triacylglycerols and Acetyl-Triacylglycerols Containing Medium-Chain Fatty Acids.” *Frontiers in Energy Research* 9. <https://doi.org/10.3389/fenrg.2021.620118>.
- Ewels, Philip A., Alexander Peltzer, Sven Fillinger, Harshil Patel, Johannes Alneberg, Andreas Wilm, Maxime Ulysse Garcia, Paolo Di Tommaso, and Sven Nahnsen. 2020. “The Nf-Core Framework for Community-Curated Bioinformatics Pipelines.” *Nature Biotechnology* 38 (3): 276–78.
- Ewing, B., and P. Green. 1998. “Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities.” *Genome Research* 8 (3): 186–94.
- Fan, Jiqing, David R. Shonnard, Tom N. Kalnes, Peter B. Johnsen, and Serin Rao. 2013. “A Life Cycle Assessment of Pennycress (*Thlaspi Arvense* L.) -Derived Jet Fuel and Diesel.” *Biomass and Bioenergy* 55 (August): 87–100.
- Feng, Hao, Karen N. Conneely, and Hao Wu. 2014. “A Bayesian Hierarchical Model to Detect Differentially Methylated Loci from Single Nucleotide Resolution Sequencing Data.” *Nucleic Acids Research* 42 (8): e69.
- Feng, Suhua, Shawn J. Cokus, Veit Schubert, Jixian Zhai, Matteo Pellegrini, and Steven E. Jacobsen. 2014. “Genome-Wide Hi-C Analyses in Wild-Type and Mutants Reveal High-Resolution Chromatin Interactions in *Arabidopsis*.” *Molecular Cell* 55 (5): 694–707.
- Feng, Suhua, Shawn J. Cokus, Xiaoyu Zhang, Pao-Yang Chen, Magnolia Bostick, Mary G. Goll, Jonathan Hetzel, et al. 2010. “Conservation and Divergence of Methylation Patterning in Plants and Animals.” *Proceedings of the National Academy of Sciences of the United States of America* 107 (19): 8689–94.
- Feng, Suhua, Steven E. Jacobsen, and Wolf Reik. 2010. “Epigenetic Reprogramming in Plant and Animal Development.” *Science* 330 (6004): 622–27.
- Fernie, Alisdair R., and Jose Gutierrez-Marcos. 2019. “From Genome to Phenome: Genome-Wide Association Studies and Other Approaches That Bridge the Genotype to Phenotype Gap.” *The Plant Journal: For Cell and Molecular Biology* 97 (1): 5–7.

- Finnegan, E. J., W. J. Peacock, and E. S. Dennis. 1996. "Reduced DNA Methylation in *Arabidopsis Thaliana* Results in Abnormal Plant Development." *Proceedings of the National Academy of Sciences of the United States of America* 93 (16): 8449–54.
- Florea, Liliana, Alexander Souvorov, Theodore S. Kalbfleisch, and Steven L. Salzberg. 2011. "Genome Assembly Has a Major Impact on Gene Content: A Comparison of Annotation in Two *Bos Taurus* Assemblies." *PloS One* 6 (6): e21400.
- Flusberg, Benjamin A., Dale R. Webster, Jessica H. Lee, Kevin J. Travers, Eric C. Olivares, Tyson A. Clark, Jonas Korlach, and Stephen W. Turner. 2010. "Direct Detection of DNA Methylation during Single-Molecule, Real-Time Sequencing." *Nature Methods* 7 (6): 461–65.
- Flynn, Jullien M., Robert Hubley, Clément Goubert, Jeb Rosen, Andrew G. Clark, Cédric Feschotte, and Arian F. Smit. 2020. "RepeatModeler2 for Automated Genomic Discovery of Transposable Element Families." *Proceedings of the National Academy of Sciences of the United States of America* 117 (17): 9451–57.
- Fojtová, M., A. Kovarik, and R. Matyásek. 2001. "Cytosine Methylation of Plastid Genome in Higher Plants. Fact or Artefact?" *Plant Science: An International Journal of Experimental Plant Biology* 160 (4): 585–93.
- Fonseca, Nuno A., Johan Rung, Alvis Brazma, and John C. Marioni. 2012. "Tools for Mapping High-Throughput Sequencing Data." *Bioinformatics* 28 (24): 3169–77.
- Franzke, Andreas, Martin A. Lysak, Ihsan A. Al-Shehbaz, Marcus A. Koch, and Klaus Mummenhoff. 2011. "Cabbage Family Affairs: The Evolutionary History of Brassicaceae." *Trends in Plant Science* 16 (2): 108–16.
- Frith, Martin C., Ryota Mori, and Kiyoshi Asai. 2012. "A Mostly Traditional Approach Improves Alignment of Bisulfite-Converted DNA." *Nucleic Acids Research* 40 (13): e100.
- 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." *Proceedings of the National Academy of Sciences of the United States of America* 89 (5): 1827–31.
- Galanti, Dario, Daniela Ramos-Cruz, Adam Nunn, Isaac Rodríguez-Arévalo, J. F. Scheepens, Claude Becker, and Oliver Bossdorf. 2022. "Genetic and Environmental Drivers of Large-Scale Epigenetic Variation in *Thlaspi Arvense*." *bioRxiv*. <https://doi.org/10.1101/2022.03.16.484610>.

Bibliography

- Gao, Shengjie, Dan Zou, Likai Mao, Huayu Liu, Pengfei Song, Youguo Chen, Shancen Zhao, et al. 2015. “BS-SNPper: SNP Calling in Bisulfite-Seq Data.” *Bioinformatics* 31 (24): 4006–8.
- Garrison, Erik, and Gabor Marth. 2012. “Haplotype-Based Variant Detection from Short-Read Sequencing.” arXiv [q-bio.GN]. arXiv. <http://arxiv.org/abs/1207.3907>.
- Gatter, Thomas, and Peter F. Stadler. 2019. “Ryūtō: Network-Flow Based Transcriptome Reconstruction.” *BMC Bioinformatics* 20 (1): 190.
- Gawehns, Fleur, Maarten Postuma, Morgane van Antro, Adam Nunn, Bernice Sepers, Samar Fatma, Thomas P. van Gurp, et al. 2022. “epiGBS2: Improvements and Evaluation of Highly Multiplexed, epiGBS-Based Reduced Representation Bisulfite Sequencing.” *Molecular Ecology Resources*, February. <https://doi.org/10.1111/1755-0998.13597>.
- Gehring, Mary, Jin Hoe Huh, Tzung-Fu Hsieh, Jon Penterman, Yeonhee Choi, John J. Harada, Robert B. Goldberg, and Robert L. Fischer. 2006. “DEMETER DNA Glycosylase Establishes MEDEA Polycomb Gene Self-Imprinting by Allele-Specific Demethylation.” *Cell* 124 (3): 495–506.
- Geng, Yupeng, Na Chang, Yuewan Zhao, Xiaoying Qin, Shugang Lu, M. James C. Crabbe, Yabin Guan, and Ticao Zhang. 2020. “Increased Epigenetic Diversity and Transient Epigenetic Memory in Response to Salinity Stress in *Thlaspi Arvense*.” *Ecology and Evolution* 10 (20): 11622–30.
- Geng, Yupeng, Yabin Guan, La Qiong, Shugang Lu, Miao An, M. James C. Crabbe, Ji Qi, Fangqing Zhao, Qin Qiao, and Ticao Zhang. 2021. “Genomic Analysis of Field Pennycress (*Thlaspi Arvense*) Provides Insights into Mechanisms of Adaptation to High Elevation.” *BMC Biology* 19 (1): 143.
- Ghurye, Jay, Mihai Pop, Sergey Koren, Derek Bickhart, and Chen-Shan Chin. 2017. “Scaffolding of Long Read Assemblies Using Long Range Contact Information.” *BMC Genomics* 18 (1): 527.
- Goodstein, David M., Shengqiang Shu, Russell Howson, Rochak Neupane, Richard D. Hayes, Joni Fazo, Therese Mitros, et al. 2012. “Phytozome: A Comparative Platform for Green Plant Genomics.” *Nucleic Acids Research* 40 (Database issue): D1178–86.
- Goodwin, Sara, John D. McPherson, and W. Richard McCombie. 2016. “Coming of Age: Ten Years of next-Generation Sequencing Technologies.” *Nature Reviews. Genetics* 17 (6): 333–51.
- Gouil, Quentin, and David C. Baulcombe. 2016. “DNA Methylation Signatures of the Plant Chromomethyltransferases.” *PLoS Genetics* 12 (12): e1006526.

- Greagg, M. A., M. J. Fogg, G. Panayotou, S. J. Evans, B. A. Connolly, and L. H. Pearl. 1999. "A Read-Ahead Function in Archaeal DNA Polymerases Detects Promutagenic Template-Strand Uracil." *Proceedings of the National Academy of Sciences of the United States of America* 96 (16): 9045–50.
- Grehl, Claudius, Markus Kuhlmann, Claude Becker, Bruno Glaser, and Ivo Grosse. 2018. "How to Design a Whole-Genome Bisulfite Sequencing Experiment." *Epigenomes* 2 (4): 21.
- Grob, Stefan, Marc W. Schmid, and Ueli Grossniklaus. 2014. "Hi-C Analysis in *Arabidopsis* Identifies the KNOT, a Structure with Similarities to the Flamenco Locus of *Drosophila*." *Molecular Cell* 55 (5): 678–93.
- Grüning, Björn, Ryan Dale, Andreas Sjödin, Brad A. Chapman, Jillian Rowe, Christopher H. Tomkins-Tinch, Renan Valieris, Johannes Köster, and Bioconda Team. 2018. "Bioconda: Sustainable and Comprehensive Software Distribution for the Life Sciences." *Nature Methods* 15 (7): 475–76.
- Guan, Dengfeng, Shane A. McCarthy, Jonathan Wood, Kerstin Howe, Yadong Wang, and Richard Durbin. 2020. "Identifying and Removing Haplotypic Duplication in Primary Genome Assemblies." *Bioinformatics* 36 (9): 2896–98.
- Gugger, Paul F., Sorel Fitz-Gibbon, Matteo PellEgrini, and Victoria L. Sork. 2016. "Species-Wide Patterns of DNA Methylation Variation in *Quercus Lobata* and Their Association with Climate Gradients." *Molecular Ecology* 25 (8): 1665–80.
- Guo, Weilong, Petko Fiziev, Weihong Yan, Shawn Cokus, Xueguang Sun, Michael Q. Zhang, Pao-Yang Chen, and Matteo Pellegrini. 2013. "BS-Seeker2: A Versatile Aligning Pipeline for Bisulfite Sequencing Data." *BMC Genomics* 14 (November): 774.
- Gurevich, Alexey, Vladislav Saveliev, Nikolay Vyahhi, and Glenn Tesler. 2013. "QUAST: Quality Assessment Tool for Genome Assemblies." *Bioinformatics* 29 (8): 1072–75.
- Gurp, Thomas P. van, Niels C. A. M. Wagemaker, Björn Wouters, Philippine Vergeer, Joop N. J. Ouborg, and Koen J. F. Verhoeven. 2016. "epiGBS: Reference-Free Reduced Representation Bisulfite Sequencing." *Nature Methods* 13 (4): 322–24.
- Hagmann, Jörg, Claude Becker, Jonas Müller, Oliver Stegle, Rhonda C. Meyer, George Wang, Korbinian Schneeberger, et al. 2015. "Century-Scale Methylome Stability in a Recently Diverged *Arabidopsis Thaliana* Lineage." *PLoS Genetics* 11 (1): e1004920.

Bibliography

- Hansen, Kasper D., Benjamin Langmead, and Rafael A. Irizarry. 2012. "BSmooth: From Whole Genome Bisulfite Sequencing Reads to Differentially Methylated Regions." *Genome Biology* 13 (10): R83.
- Hardcastle, Thomas J., Sebastian Y. Müller, and David C. Baulcombe. 2018. "Towards Annotating the Plant Epigenome: The Arabidopsis Thaliana Small RNA Locus Map." *Scientific Reports* 8 (1): 6338.
- Hattman, S., S. Schlagman, and L. Cousens. 1973. "Isolation of a Mutant of Escherichia Coli Defective in Cytosine-Specific Deoxyribonucleic Acid Methylase Activity and in Partial Protection of Bacteriophage Lambda against Restriction by Cells Containing the N-3 Drug-Resistance Factor." *Journal of Bacteriology* 115 (3): 1103–7.
- Hayatsu, H., and M. Shiragami. 1979. "Reaction of Bisulfite with the 5-Hydroxymethyl Group in Pyrimidines and in Phage DNAs." *Biochemistry* 18 (4): 632–37.
- Hayatsu, H., Y. Wataya, K. Kai, and S. Iida. 1970. "Reaction of Sodium Bisulfite with Uracil, Cytosine, and Their Derivatives." *Biochemistry* 9 (14): 2858–65.
- Heard, Edith, and Robert A. Martienssen. 2014. "Transgenerational Epigenetic Inheritance: Myths and Mechanisms." *Cell* 157 (1): 95–109.
- Heer, Katrin, Jeannie Mounger, M. Teresa Boquete, Christina L. Richards, and Lars Opgenoorth. 2018. "The Diversifying Field of Plant Epigenetics." *The New Phytologist* 217 (3): 988–92.
- He, Guangming, Beibei Chen, Xuncheng Wang, Xueyong Li, Jigang Li, Hang He, Mei Yang, et al. 2013. "Conservation and Divergence of Transcriptomic and Epigenomic Variation in Maize Hybrids." *Genome Biology* 14 (6): R57.
- Henderson, Ian R., and Steven E. Jacobsen. 2007. "Epigenetic Inheritance in Plants." *Nature* 447 (7143): 418–24.
- Heyn, Holger, Sebastian Moran, Irene Hernando-Herraez, Sergi Sayols, Antonio Gomez, Juan Sandoval, Dave Monk, et al. 2013. "DNA Methylation Contributes to Natural Human Variation." *Genome Research* 23 (9): 1363–72.
- Hodges, Emily, Andrew D. Smith, Jude Kendall, Zhenyu Xuan, Kandasamy Ravi, Michelle Rooks, Michael Q. Zhang, et al. 2009. "High Definition Profiling of Mammalian DNA Methylation by Array Capture and Single Molecule Bisulfite Sequencing." *Genome Research* 19 (9): 1593–1605.

- Hofmann, Nancy R. 2015. "Epigenetic Battles Underfoot: Allelopathy among Plants Can Target Chromatin Modification." *The Plant Cell* 27 (11): 3021.
- Huang, Huan, Ruie Liu, Qingfeng Niu, Kai Tang, Bo Zhang, Heng Zhang, Kunsong Chen, Jian-Kang Zhu, and Zhaobo Lang. 2019. "Global Increase in DNA Methylation during Orange Fruit Development and Ripening." *Proceedings of the National Academy of Sciences of the United States of America* 116 (4): 1430–36.
- Huang, Xuehui, Xinghua Wei, Tao Sang, Qiang Zhao, Qi Feng, Yan Zhao, Canyang Li, et al. 2010. "Genome-Wide Association Studies of 14 Agronomic Traits in Rice Landraces." *Nature Genetics* 42 (11): 961–67.
- Huber, Sabrina M., Pieter van Delft, Lee Mendil, Martin Bachman, Katherine Smollett, Finn Werner, Eric A. Miska, and Shankar Balasubramanian. 2015. "Formation and Abundance of 5-Hydroxymethylcytosine in RNA." *Chembiochem: A European Journal of Chemical Biology* 16 (5): 752–55.
- Huff, Jason T., and Daniel Zilberman. 2014. "Dnmt1-Independent CG Methylation Contributes to Nucleosome Positioning in Diverse Eukaryotes." *Cell* 156 (6): 1286–97.
- Hughes, A. Randall, Brian D. Inouye, Marc T. J. Johnson, Nora Underwood, and Mark Vellend. 2008. "Ecological Consequences of Genetic Diversity." *Ecology Letters* 11 (6): 609–23.
- Hwang, Eun-Young, Qijian Song, Gaofeng Jia, James E. Specht, David L. Hyten, Jose Costa, and Perry B. Cregan. 2014. "A Genome-Wide Association Study of Seed Protein and Oil Content in Soybean." *BMC Genomics* 15 (January): 1.
- Jackson, James P., Anders M. Lindroth, Xiaofeng Cao, and Steven E. Jacobsen. 2002. "Control of CpNpG DNA Methylation by the KRYPTONITE Histone H3 Methyltransferase." *Nature* 416 (6880): 556–60.
- Jaenisch, Rudolf, and Adrian Bird. 2003. "Epigenetic Regulation of Gene Expression: How the Genome Integrates Intrinsic and Environmental Signals." *Nature Genetics* 33 Suppl (March): 245–54.
- Jarvis, Brice A., Trevor B. Romsdahl, Michaela G. McGinn, Tara J. Nazareus, Edgar B. Cahoon, Kent D. Chapman, and John C. Sedbrook. 2021. "CRISPR/Cas9-Induced *fad2* and *rod1* Mutations Stacked With *fae1* Confer High Oleic Acid Seed Oil in Pennycress (*Thlaspi Arvense* L.)." *Frontiers in Plant Science* 12 (April): 652319.

Bibliography

- Jean Finnegan, E., Kathryn A. Kovac, Estelle Jaligot, Candice C. Sheldon, W. James Peacock, and Elizabeth S. Dennis. 2005. "The Downregulation of FLOWERING LOCUS C (FLC) Expression in Plants with Low Levels of DNA Methylation and by Vernalization Occurs by Distinct Mechanisms." *The Plant Journal: For Cell and Molecular Biology* 44 (3): 420–32.
- Jeddeloh, J. A., T. L. Stokes, and E. J. Richards. 1999. "Maintenance of Genomic Methylation Requires a SWI2/SNF2-like Protein." *Nature Genetics* 22 (1): 94–97.
- Jeltsch, Albert. 2006. "On the Enzymatic Properties of Dnmt1: Specificity, Processivity, Mechanism of Linear Diffusion and Allosteric Regulation of the Enzyme." *Epigenetics: Official Journal of the DNA Methylation Society* 1 (2): 63–66.
- Jez, Joseph M., Christopher N. Topp, M. David Marks, Ratan Chopra, and John C. Sedbrook. 2021. "Technologies Enabling Rapid Crop Improvements for Sustainable Agriculture: Example Pennycress (*Thlaspi Arvense* L.)." *Emerging Topics in Life Sciences* 5 (2): 325–35.
- Johannes, Frank, Emmanuelle Porcher, Felipe K. Teixeira, Vera Saliba-Colombani, Matthieu Simon, Nicolas Agier, Agnès Bulski, et al. 2009. "Assessing the Impact of Transgenerational Epigenetic Variation on Complex Traits." *PLoS Genetics* 5 (6): e1000530.
- Johannes, Frank, and Robert J. Schmitz. 2019. "Spontaneous Epimutations in Plants." *The New Phytologist* 221 (3): 1253–59.
- Johnson, Gregg A., Michael B. Kantar, Kevin J. Betts, and Donald L. Wyse. 2015. "Field Pennycress Production and Weed Control in a Double Crop System with Soybean in Minnesota." *Agronomy Journal* 107 (2): 532–40.
- Jones, Philip, David Binns, Hsin-Yu Chang, Matthew Fraser, Weizhong Li, Craig McAnulla, Hamish McWilliam, et al. 2014. "InterProScan 5: Genome-Scale Protein Function Classification." *Bioinformatics* 30 (9): 1236–40.
- Jühling, Frank, Helene Kretzmer, Stephan H. Bernhart, Christian Otto, Peter F. Stadler, and Steve Hoffmann. 2016. "Metilene: Fast and Sensitive Calling of Differentially Methylated Regions from Bisulfite Sequencing Data." *Genome Research* 26 (2): 256–62.
- Jullien, Pauline E., Assaf Mosquana, Mathieu Ingouff, Tadashi Sakata, Nir Ohad, and Frédéric Berger. 2008. "Retinoblastoma and Its Binding Partner MSI1 Control Imprinting in Arabidopsis." *PLoS Biology* 6 (8): e194.

- Kankel, Mark W., Douglas E. Ramsey, Trevor L. Stokes, Susan K. Flowers, Jeremy R. Haag, Jeffrey A. Jeddelloh, Nicole C. Riddle, Michelle L. Verbsky, and Eric J. Richards. 2003. "Arabidopsis MET1 Cytosine Methyltransferase Mutants." *Genetics* 163 (3): 1109–22.
- Karimzadeh, Mehran, Carl Ernst, Anshul Kundaje, and Michael M. Hoffman. 2018. "Umap and Bimap: Quantifying Genome and Methylome Mappability." *Nucleic Acids Research* 46 (20): e120.
- Kawakatsu, Taiji, Shao-Shan Carol Huang, Florian Jupe, Eriko Sasaki, Robert J. Schmitz, Mark A. Urich, Rosa Castanon, et al. 2016. "Epigenomic Diversity in a Global Collection of Arabidopsis Thaliana Accessions." *Cell* 166 (2): 492–505.
- Kawashima, Tomokazu, and Frédéric Berger. 2014. "Epigenetic Reprogramming in Plant Sexual Reproduction." *Nature Reviews. Genetics* 15 (9): 613–24.
- Kelly, Scott A., Tami M. Panhuis, and Andrew M. Stoehr. 2012. "Phenotypic Plasticity: Molecular Mechanisms and Adaptive Significance." *Comprehensive Physiology* 2 (2): 1417–39.
- Kenchanmane Raju, Sunil K., Eleanore Jeanne Ritter, and Chad E. Niederhuth. 2019. "Establishment, Maintenance, and Biological Roles of Non-CG Methylation in Plants." *Essays in Biochemistry* 63 (6): 743–55.
- Kerkel, Kristi, Alexandra Spadola, Eric Yuan, Jolanta Kosek, Le Jiang, Eldad Hod, Kerry Li, et al. 2008. "Genomic Surveys by Methylation-Sensitive SNP Analysis Identify Sequence-Dependent Allele-Specific DNA Methylation." *Nature Genetics* 40 (7): 904–8.
- Kim, Jeongsik, Jin Hee Kim, Eric J. Richards, Kyung Min Chung, and Hye Ryun Woo. 2014. "Arabidopsis VIM Proteins Regulate Epigenetic Silencing by Modulating DNA Methylation and Histone Modification in Cooperation with MET1." *Molecular Plant* 7 (9): 1470–85.
- Kinoshita, Tetsu, Asuka Miura, Yeonhee Choi, Yuki Kinoshita, Xiaofeng Cao, Steven E. Jacobsen, Robert L. Fischer, and Tetsuji Kakutani. 2004. "One-Way Control of FWA Imprinting in Arabidopsis Endosperm by DNA Methylation." *Science* 303 (5657): 521–23.
- Kinoshita, Tetsu, and Motoaki Seki. 2014. "Epigenetic Memory for Stress Response and Adaptation in Plants." *Plant & Cell Physiology* 55 (11): 1859–63.
- Kint, Sam, Ward De Spiegelaere, Jonas De Kesel, Linos Vandekerckhove, and Wim Van Criekinge. 2018. "Evaluation of Bisulfite Kits for DNA Methylation Profiling in Terms of DNA Fragmentation and DNA Recovery Using Digital PCR." *PloS One* 13 (6): e0199091.

Bibliography

- Koboldt, Daniel C., Ken Chen, Todd Wylie, David E. Larson, Michael D. McLellan, Elaine R. Mardis, George M. Weinstock, Richard K. Wilson, and Li Ding. 2009. "VarScan: Variant Detection in Massively Parallel Sequencing of Individual and Pooled Samples." *Bioinformatics* 25 (17): 2283–85.
- Köhler, Claudia, Damian R. Page, Valeria Gagliardini, and Ueli Grossniklaus. 2005. "The Arabidopsis Thaliana MEDEA Polycomb Group Protein Controls Expression of PHERES1 by Parental Imprinting." *Nature Genetics* 37 (1): 28–30.
- Koren, Sergey, Brian P. Walenz, Konstantin Berlin, Jason R. Miller, Nicholas H. Bergman, and Adam M. Phillippy. 2017. "Canu: Scalable and Accurate Long-Read Assembly via Adaptive K-Mer Weighting and Repeat Separation." *Genome Research* 27 (5): 722–36.
- Korf, Ian. 2004. "Gene Finding in Novel Genomes." *BMC Bioinformatics* 5 (May): 59.
- Korthauer, Keegan, Sutirtha Chakraborty, Yuval Benjamini, and Rafael A. Irizarry. 2018. "Detection and Accurate False Discovery Rate Control of Differentially Methylated Regions from Whole Genome Bisulfite Sequencing." *Biostatistics* 20 (3): 367–83.
- Köster, Johannes, and Sven Rahmann. 2012. "Snakemake—a Scalable Bioinformatics Workflow Engine." *Bioinformatics* 28 (19): 2520–22.
- Kozarewa, Iwanka, Zemin Ning, Michael A. Quail, Mandy J. Sanders, Matthew Berriman, and Daniel J. Turner. 2009. "Amplification-Free Illumina Sequencing-Library Preparation Facilitates Improved Mapping and Assembly of (G+C)-Biased Genomes." *Nature Methods* 6 (4): 291–95.
- Kreutz, Clemens, Nilay S. Can, Ralf Schulze Bruening, Rabea Meyberg, Zsuzsanna Mérai, Noe Fernandez-Pozo, and Stefan A. Rensing. 2020. "A Blind and Independent Benchmark Study for Detecting Differentially Methylated Regions in Plants." *Bioinformatics* 36 (17): 4673.
- Krueger, Felix, and Simon R. Andrews. 2011. "Bismark: A Flexible Aligner and Methylation Caller for Bisulfite-Seq Applications." *Bioinformatics* 27 (11): 1571–72.
- Krzywinski, Martin, Jacqueline Schein, Inanç Birol, Joseph Connors, Randy Gascoyne, Doug Horsman, Steven J. Jones, and Marco A. Marra. 2009. "Circos: An Information Aesthetic for Comparative Genomics." *Genome Research* 19 (9): 1639–45.
- Kunde-Ramamoorthy, Govindarajan, Cristian Coarfa, Eleonora Laritsky, Noah J. Kessler, R. Alan Harris, Mingchu Xu, Rui Chen, Lanlan Shen, Aleksandar Milosavljevic, and Robert A. Waterland.

2014. “Comparison and Quantitative Verification of Mapping Algorithms for Whole-Genome Bisulfite Sequencing.” *Nucleic Acids Research* 42 (6): e43.
- Kurtzer, Gregory M., Vanessa Sochat, and Michael W. Bauer. 2017. “Singularity: Scientific Containers for Mobility of Compute.” *PloS One* 12 (5): e0177459.
- Lagesen, Karin, Peter Hallin, Einar Andreas Rødland, Hans-Henrik Staerfeldt, Torbjørn Rognes, and David W. Ussery. 2007. “RNAmmer: Consistent and Rapid Annotation of Ribosomal RNA Genes.” *Nucleic Acids Research* 35 (9): 3100–3108.
- La, Honggui, Bo Ding, Gyan P. Mishra, Bo Zhou, Hongmei Yang, Maria del Rosario Bellizzi, Songbiao Chen, et al. 2011. “A 5-Methylcytosine DNA Glycosylase/lyase Demethylates the Retrotransposon Tos17 and Promotes Its Transposition in Rice.” *Proceedings of the National Academy of Sciences of the United States of America* 108 (37): 15498–503.
- Lam, Ernest T., Alex Hastie, Chin Lin, Dean Ehrlich, Somes K. Das, Michael D. Austin, Paru Deshpande, et al. 2012. “Genome Mapping on Nanochannel Arrays for Structural Variation Analysis and Sequence Assembly.” *Nature Biotechnology* 30 (8): 771–76.
- Langmead, Ben, and Steven L. Salzberg. 2012. “Fast Gapped-Read Alignment with Bowtie 2.” *Nature Methods* 9 (4): 357–59.
- Lang-Mladek, Christina, Olga Popova, Kathrin Kiok, Marc Berlinger, Branislava Rakic, Werner Aufsatz, Claudia Jonak, Marie-Theres Hauser, and Christian Luschnig. 2010. “Transgenerational Inheritance and Resetting of Stress-Induced Loss of Epigenetic Gene Silencing in Arabidopsis.” *Molecular Plant* 3 (3): 594–602.
- Lang, Zhaobo, Yihai Wang, Kai Tang, Dengguo Tang, Tatsiana Datsenka, Jingfei Cheng, Yijing Zhang, Avtar K. Handa, and Jian-Kang Zhu. 2017. “Critical Roles of DNA Demethylation in the Activation of Ripening-Induced Genes and Inhibition of Ripening-Repressed Genes in Tomato Fruit.” *Proceedings of the National Academy of Sciences of the United States of America* 114 (22): E4511–19.
- Lappalainen, Tuuli, and John M. Greally. 2017. “Associating Cellular Epigenetic Models with Human Phenotypes.” *Nature Reviews. Genetics* 18 (7): 441–51.
- Latzel, Vít, Eric Allan, Amanda Bortolini Silveira, Vincent Colot, Markus Fischer, and Oliver Bossdorf. 2013. “Epigenetic Diversity Increases the Productivity and Stability of Plant Populations.” *Nature Communications* 4: 2875.

Bibliography

- Law, Julie A., and Steven E. Jacobsen. 2010. "Establishing, Maintaining and Modifying DNA Methylation Patterns in Plants and Animals." *Nature Reviews. Genetics* 11 (3): 204–20.
- Lei, Mingguang, Huiming Zhang, Russell Julian, Kai Tang, Shaojun Xie, and Jian-Kang Zhu. 2015. "Regulatory Link between DNA Methylation and Active Demethylation in Arabidopsis." *Proceedings of the National Academy of Sciences of the United States of America* 112 (11): 3553–57.
- Leutwiler, Leslie S., Barbara R. Hough-Evans, and Elliot M. Meyerowitz. 1984. "The DNA of Arabidopsis Thaliana." *Molecular & General Genetics: MGG* 194 (1): 15–23.
- Levy, Shawn E., and Braden E. Boone. 2019. "Next-Generation Sequencing Strategies." *Cold Spring Harbor Perspectives in Medicine* 9 (7). <https://doi.org/10.1101/cshperspect.a025791>.
- Liao, Yang, Gordon K. Smyth, and Wei Shi. 2014. "featureCounts: An Efficient General Purpose Program for Assigning Sequence Reads to Genomic Features." *Bioinformatics* 30 (7): 923–30.
- Li, E., C. Beard, A. C. Forster, T. H. Bestor, and R. Jaenisch. 1993. "DNA Methylation, Genomic Imprinting, and Mammalian Development." *Cold Spring Harbor Symposia on Quantitative Biology* 58: 297–305.
- Liégard, Benjamin, Victoire Baillet, Mathilde Etcheverry, Evens Joseph, Christine Lariagon, Jocelyne Lemoine, Aurélie Evrard, et al. 2019. "Quantitative Resistance to Clubroot Infection Mediated by Transgenerational Epigenetic Variation in Arabidopsis." *The New Phytologist* 222 (1): 468–79.
- Li, Fay-Wei, and Alex Harkess. 2018. "A Guide to Sequence Your Favorite Plant Genomes." *Applications in Plant Sciences* 6 (3): e1030.
- Li, Heng. 2011. "A Statistical Framework for SNP Calling, Mutation Discovery, Association Mapping and Population Genetical Parameter Estimation from Sequencing Data." *Bioinformatics* 27 (21): 2987–93.
- . 2018. "Minimap2: Pairwise Alignment for Nucleotide Sequences." *Bioinformatics* 34 (18): 3094–3100.
- Li, Heng, and Richard Durbin. 2009. "Fast and Accurate Short Read Alignment with Burrows–Wheeler Transform." *Bioinformatics* 25 (14): 1754–60.

- Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. "The Sequence Alignment/Map Format and SAMtools." *Bioinformatics* 25 (16): 2078–79.
- Li, Heng, and Nils Homer. 2010. "A Survey of Sequence Alignment Algorithms for next-Generation Sequencing." *Briefings in Bioinformatics* 11 (5): 473–83.
- Lindahl, T., and R. D. Wood. 1999. "Quality Control by DNA Repair." *Science* 286 (5446): 1897–1905.
- Lin, Tao, Guangtao Zhu, Junhong Zhang, Xiangyang Xu, Qinghui Yu, Zheng Zheng, Zhonghua Zhang, et al. 2014. "Genomic Analyses Provide Insights into the History of Tomato Breeding." *Nature Genetics* 46 (11): 1220–26.
- Lin, Xueqiu, Deqiang Sun, Benjamin Rodriguez, Qian Zhao, Hanfei Sun, Yong Zhang, and Wei Li. 2013. "BSeQC: Quality Control of Bisulfite Sequencing Experiments." *Bioinformatics* 29 (24): 3227–29.
- Lippman, Zachary, Anne-Valérie Gendrel, Michael Black, Matthew W. Vaughn, Neilay Dedhia, W. Richard McCombie, Kimberly Lavine, et al. 2004. "Role of Transposable Elements in Heterochromatin and Epigenetic Control." *Nature* 430 (6998): 471–76.
- 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." *Cell* 133 (3): 523–36.
- Lister, Ryan, Mattia Pelizzola, Robert H. Downen, R. David Hawkins, Gary Hon, Julian Tonti-Filippini, Joseph R. Nery, et al. 2009. "Human DNA Methylomes at Base Resolution Show Widespread Epigenomic Differences." *Nature* 462 (7271): 315–22.
- Liu, Yaping, Kimberly D. Siegmund, Peter W. Laird, and Benjamin P. Berman. 2012. "Bis-SNP: Combined DNA Methylation and SNP Calling for Bisulfite-Seq Data." *Genome Biology* 13 (7): R61.
- Liu, Yibin, Paulina Siejka-Zielińska, Gergana Velikova, Ying Bi, Fang Yuan, Marketa Tomkova, Chunsen Bai, Lei Chen, Benjamin Schuster-Böckler, and Chun-Xiao Song. 2019. "Bisulfite-Free Direct Detection of 5-Methylcytosine and 5-Hydroxymethylcytosine at Base Resolution." *Nature Biotechnology* 37 (4): 424–29.

Bibliography

- Li, Weitao, Ziwei Zhu, Mawsheng Chern, Junjie Yin, Chao Yang, Li Ran, Mengping Cheng, et al. 2017. “A Natural Allele of a Transcription Factor in Rice Confers Broad-Spectrum Blast Resistance.” *Cell* 170 (1): 114–26.e15.
- Li, Xueqin, C. Jake Harris, Zhenhui Zhong, Wei Chen, Rui Liu, Bei Jia, Zonghua Wang, Sisi Li, Steven E. Jacobsen, and Jiamu Du. 2018. “Mechanistic Insights into Plant SUVH Family H3K9 Methyltransferases and Their Binding to Context-Biased Non-CG DNA Methylation.” *Proceedings of the National Academy of Sciences of the United States of America* 115 (37): E8793–8802.
- Li, Zhenyu, Yanxiang Chen, Desheng Mu, Jianying Yuan, Yujian Shi, Hao Zhang, Jun Gan, et al. 2011. “Comparison of the Two Major Classes of Assembly Algorithms: Overlap–layout–consensus and de-Bruijn-Graph.” *Briefings in Functional Genomics* 11 (1): 25–37.
- Lloyd, James P. B., and Ryan Lister. 2022. “Epigenome Plasticity in Plants.” *Nature Reviews. Genetics* 23 (1): 55–68.
- Lomsadze, Alexandre, Vardges Ter-Hovhannisyan, Yury O. Chernoff, and Mark Borodovsky. 2005. “Gene Identification in Novel Eukaryotic Genomes by Self-Training Algorithm.” *Nucleic Acids Research* 33 (20): 6494–6506.
- Lowe, T. M., and S. R. Eddy. 1997. “tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence.” *Nucleic Acids Research* 25 (5): 955–64.
- Lunardon, Alice, Nathan R. Johnson, Emily Hagerott, Tamia Phifer, Seth Polydore, Ceyda Coruh, and Michael J. Axtell. 2020. “Integrated Annotations and Analyses of Small RNA–producing Loci from 47 Diverse Plants.” *Genome Research* 30 (3): 497–513.
- Magoč, Tanja, and Steven L. Salzberg. 2011. “FLASH: Fast Length Adjustment of Short Reads to Improve Genome Assemblies.” *Bioinformatics* 27 (21): 2957–63.
- Mahmood, Asaad M., and Jim M. Dunwell. 2019. “Evidence for Novel Epigenetic Marks within Plants.” *AIMS Genetics* 6 (4): 70–87.
- Manning, Kenneth, Mahmut Tör, Mervin Poole, Yiguo Hong, Andrew J. Thompson, Graham J. King, James J. Giovannoni, and Graham B. Seymour. 2006. “A Naturally Occurring Epigenetic Mutation in a Gene Encoding an SBP-Box Transcription Factor Inhibits Tomato Fruit Ripening.” *Nature Genetics* 38 (8): 948–52.

- Manolio, Teri A., Francis S. Collins, Nancy J. Cox, David B. Goldstein, Lucia A. Hindorff, David J. Hunter, Mark I. McCarthy, et al. 2009. "Finding the Missing Heritability of Complex Diseases." *Nature* 461 (7265): 747–53.
- Marano, M. R., and N. Carrillo. 1991. "Chromoplast Formation during Tomato Fruit Ripening. No Evidence for Plastid DNA Methylation." *Plant Molecular Biology* 16 (1): 11–19.
- Marçais, Guillaume, Arthur L. Delcher, Adam M. Phillippy, Rachel Coston, Steven L. Salzberg, and Aleksey Zimin. 2018. "MUMmer4: A Fast and Versatile Genome Alignment System." *PLoS Computational Biology* 14 (1): e1005944.
- Marçais, Guillaume, and Carl Kingsford. 2011. "A Fast, Lock-Free Approach for Efficient Parallel Counting of Occurrences of K-Mers." *Bioinformatics* 27 (6): 764–70.
- Marco-Sola, Santiago, Michael Sammeth, Roderic Guigó, and Paolo Ribeca. 2012. "The GEM Mapper: Fast, Accurate and Versatile Alignment by Filtration." *Nature Methods* 9 (12): 1185–88.
- Mardis, Elaine, John McPherson, Robert Martienssen, Richard K. Wilson, and W. Richard McCombie. 2002. "What Is Finished, and Why Does It Matter." *Genome Research* 12 (5): 669–71.
- Margueron, Raphaël, and Danny Reinberg. 2010. "Chromatin Structure and the Inheritance of Epigenetic Information." *Nature Reviews. Genetics* 11 (4): 285–96.
- Martin, Antoine, Christelle Troadec, Adnane Boualem, Mazen Rajab, Ronan Fernandez, Halima Morin, Michel Pitrat, Catherine Dogimont, and Abdelhafid Bendahmane. 2009. "A Transposon-Induced Epigenetic Change Leads to Sex Determination in Melon." *Nature* 461 (7267): 1135–38.
- Martin, Marcel. 2011. "Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads." *EMBnet.journal* 17 (1): 10–12.
- Martin, Marcel, Murray Patterson, Shilpa Garg, Sarah O. Fischer, Nadia Pisanti, Gunnar W. Klau, Alexander Schöenhuth, and Tobias Marschall. 2016. "WhatsHap: Fast and Accurate Read-Based Phasing." *bioRxiv*. <https://doi.org/10.1101/085050>.
- Matzke, Marjori A., and Rebecca A. Mosher. 2014. "RNA-Directed DNA Methylation: An Epigenetic Pathway of Increasing Complexity." *Nature Reviews. Genetics* 15 (6): 394–408.

Bibliography

McGinn, Michaela, Winthrop B. Phippen, Ratan Chopra, Sunil Bansal, Brice A. Jarvis, Mary E. Phippen, Kevin M. Dorn, et al. 2019. "Molecular Tools Enabling Pennycress (*Thlaspi Arvense*) as a Model Plant and Oilseed Cash Cover Crop." *Plant Biotechnology Journal* 17 (4): 776–88.

McKenna, Aaron, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, et al. 2010. "The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data." *Genome Research* 20 (9): 1297–1303.

McKernan, Kevin Judd, Heather E. Peckham, Gina L. Costa, Stephen F. McLaughlin, Yutao Fu, Eric F. Tsung, Christopher R. Clouser, et al. 2009. "Sequence and Structural Variation in a Human Genome Uncovered by Short-Read, Massively Parallel Ligation Sequencing Using Two-Base Encoding." *Genome Research* 19 (9): 1527–41.

Medrano, Mónica, Carlos M. Herrera, and Pilar Bazaga. 2014. "Epigenetic Variation Predicts Regional and Local Intraspecific Functional Diversity in a Perennial Herb." *Molecular Ecology* 23 (20): 4926–38.

Mehrotra, Shweta, and Vinod Goyal. 2014. "Repetitive Sequences in Plant Nuclear DNA: Types, Distribution, Evolution and Function." *Genomics, Proteomics & Bioinformatics* 12 (4): 164–71.

Meinke, D. W., J. M. Cherry, C. Dean, S. D. Rounsley, and M. Koornneef. 1998. "Arabidopsis Thaliana: A Model Plant for Genome Analysis." *Science* 282 (5389): 662, 679–82.

Meissner, Alexander, Andreas Gnirke, George W. Bell, Bernard Ramsahoye, Eric S. Lander, and Rudolf Jaenisch. 2005. "Reduced Representation Bisulfite Sequencing for Comparative High-Resolution DNA Methylation Analysis." *Nucleic Acids Research* 33 (18): 5868–77.

Merkel. 2014. "Docker: Lightweight Linux Containers for Consistent Development and Deployment." *Linux Journal*.
<https://www.seltzer.com/margo/teaching/CS508.19/papers/merkel14.pdf>.

Merkel, Angelika, Marcos Fernández-Callejo, Eloi Casals, Santiago Marco-Sola, Ronald Schuyler, Ivo G. Gut, and Simon C. Heath. 2019. "gemBS: High Throughput Processing for DNA Methylation Data from Bisulfite Sequencing." *Bioinformatics* 35 (5): 737–42.

Metzker. 2010. "Sequencing Technologies—the next Generation." *Nature Reviews. Genetics*.
<https://www.nature.com/articles/nrg2626/boxes/bx1>.

Meyer, Peter. 2011. "DNA Methylation Systems and Targets in Plants." *FEBS Letters* 585 (13): 2008–15.

- Michalovova, M., B. Vyskot, and E. Kejnovsky. 2013. "Analysis of Plastid and Mitochondrial DNA Insertions in the Nucleus (NUPTs and NUMTs) of Six Plant Species: Size, Relative Age and Chromosomal Localization." *Heredity* 111 (4): 314–20.
- Mikheyev, Alexander S., and Mandy M. Y. Tin. 2014. "A First Look at the Oxford Nanopore MinION Sequencer." *Molecular Ecology Resources* 14 (6): 1097–1102.
- Mirouze, Marie, Jon Reinders, Etienne Bucher, Taisuke Nishimura, Korbinian Schneeberger, Stephan Ossowski, Jun Cao, Detlef Weigel, Jerzy Paszkowski, and Olivier Mathieu. 2009. "Selective Epigenetic Control of Retrotransposition in Arabidopsis." *Nature* 461 (7262): 427–30.
- Miura, A., S. Yonebayashi, K. Watanabe, T. Toyama, H. Shimada, and T. Kakutani. 2001. "Mobilization of Transposons by a Mutation Abolishing Full DNA Methylation in Arabidopsis." *Nature* 411 (6834): 212–14.
- Miura, Fumihito, Yusuke Enomoto, Ryo Dairiki, and Takashi Ito. 2012. "Amplification-Free Whole-Genome Bisulfite Sequencing by Post-Bisulfite Adaptor Tagging." *Nucleic Acids Research* 40 (17): e136.
- 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." *Proceedings of the National Academy of Sciences of the United States of America* 106 (27): 11218–23.
- 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." *Cell* 146 (6): 1029–41.
- Monroe, J. Grey, Thanvi Srikant, Pablo Carbonell-Bejerano, Claude Becker, Mariele Lensink, Moises Exposito-Alonso, Marie Klein, et al. 2022. "Mutation Bias Reflects Natural Selection in Arabidopsis Thaliana." *Nature* 602 (7895): 101–5.
- Moore, Sarah A., M. Scott Wells, Russ W. Gesch, Roger L. Becker, Carl J. Rosen, and Melissa L. Wilson. 2020. "Pennycress as a Cash Cover-Crop: Improving the Sustainability of Sweet Corn Production Systems." *Agronomy* 10 (5): 614.
- Moser, Bryan R. 2012. "Biodiesel from Alternative Oilseed Feedstocks: Camelina and Field Pennycress." *Biofuels* 3 (2): 193–209.

Bibliography

- Moser, Bryan R., Gerhard Knothe, Steven F. Vaughn, and Terry A. Isbell. 2009. "Production and Evaluation of Biodiesel from Field Pennycress (*Thlaspi Arvense* L.) Oil." *Energy & Fuels: An American Chemical Society Journal* 23 (8): 4149–55.
- Mulligan, Gerald A. 1957. "Chromosome Numbers of Canadian Weeds. I." *Canadian Journal of Botany. Journal Canadien de Botanique* 35 (5): 779–89.
- Mulligan, Gerald A., and Peter G. Kevan. 1973. "Color, Brightness, and Other Floral Characteristics Attracting Insects to the Blossoms of Some Canadian Weeds." *Canadian Journal of Botany. Journal Canadien de Botanique* 51 (10): 1939–52.
- Murat, Florent, Alexandra Louis, Florian Maumus, Alix Armero, Richard Cooke, Hadi Quesneville, Hugues Roest Crollius, and Jerome Salse. 2015. "Understanding Brassicaceae Evolution through Ancestral Genome Reconstruction." *Genome Biology* 16 (December): 262.
- Nawrocki, Eric P., and Sean R. Eddy. 2013. "Infernal 1.1: 100-Fold Faster RNA Homology Searches." *Bioinformatics* 29 (22): 2933–35.
- Needleman, S. B., and C. D. Wunsch. 1970. "A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins." *Journal of Molecular Biology* 48 (3): 443–53.
- Neumann, K., B. Kobiljski, S. Denčić, R. K. Varshney, and A. Börner. 2011. "Genome-Wide Association Mapping: A Case Study in Bread Wheat (*Triticum Aestivum* L.)." *Molecular Breeding: New Strategies in Plant Improvement* 27 (1): 37–58.
- Niederhuth, Chad E., Adam J. Bewick, Lexiang Ji, Magdy S. Alabady, Kyung Do Kim, Qing Li, Nicholas A. Rohr, et al. 2016. "Widespread Natural Variation of DNA Methylation within Angiosperms." *Genome Biology* 17 (1): 194.
- Niknafs, Yashar S., Balaji Pandian, Hariharan K. Iyer, Arul M. Chinnaiyan, and Matthew K. Iyer. 2017. "TACO Produces Robust Multisample Transcriptome Assemblies from RNA-Seq." *Nature Methods* 14 (1): 68–70.
- 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." *Nature Plants* 6 (8): 942–56.
- Nour-Eldin, Hussam Hassan, Tonni Grube Andersen, Meike Burow, Svend Roesen Madsen, Morten Egevang Jørgensen, Carl Erik Olsen, Ingo Dreyer, Rainer Hedrich, Dietmar Geiger, and

- Barbara Ann Halkier. 2012. "NRT/PTR Transporters Are Essential for Translocation of Glucosinolate Defence Compounds to Seeds." *Nature* 488 (7412): 531–34.
- Nunn, Adam, Sultan Nilay Can, Christian Otto, Mario Fasold, Bárbara Díez Rodríguez, Noé Fernández-Pozo, Stefan A. Rensing, Peter F. Stadler, and David Langenberger. 2021. "EpiDiverse Toolkit: A Pipeline Suite for the Analysis of Bisulfite Sequencing Data in Ecological Plant Epigenetics." *NAR Genomics and Bioinformatics* 3 (4): lqab106.
- Nunn, Adam, Isaac Rodríguez-Arévalo, Zenith Tandukar, Katherine Frels, Adrián Contreras-Garrido, Pablo Carbonell-Bejerano, Panpan Zhang, et al. 2022. "Chromosome-Level *Thlaspi Arvense* Genome Provides New Tools for Translational Research and for a Newly Domesticated Cash Cover Crop of the Cooler Climates." *Plant Biotechnology Journal* 20 (5): 944–63.
- 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." *Genome Biology* 19 (1): 33.
- Ong-Abdullah, Meilina, Jared M. Ordway, Nan Jiang, Siew-Eng Ooi, Sau-Yee Kok, Norashikin Sarpan, Nuraziyan Azimi, et al. 2015. "Loss of Karma Transposon Methylation Underlies the Mantled Somaclonal Variant of Oil Palm." *Nature* 525 (7570): 533–37.
- Onishi-Seebacher, Megumi, and Jan O. Korb. 2011. "Challenges in Studying Genomic Structural Variant Formation Mechanisms: The Short-Read Dilemma and beyond." *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology* 33 (11): 840–50.
- Ott, Matthew A., Carrie A. Eberle, Matt D. Thom, David W. Archer, Frank Forcella, Russell W. Gesch, and Donald L. Wyse. 2019. "Economics and Agronomics of Relay-cropping Pennycress and Camelina with Soybean in Minnesota." *Agronomy Journal* 111 (3): 1281–92.
- Otto, Christian, Peter F. Stadler, and Steve Hoffmann. 2012. "Fast and Sensitive Mapping of Bisulfite-Treated Sequencing Data." *Bioinformatics* 28 (13): 1698–1704.
- . 2014. "Lacking Alignments? The next-Generation Sequencing Mapper Segemehl Revisited." *Bioinformatics* 30 (13): 1837–43.
- Ou, Shujun, Weija Su, Yi Liao, Kapeel Chougule, Jireh R. A. Agda, Adam J. Hellinga, Carlos Santiago Blanco Lugo, et al. 2019. "Benchmarking Transposable Element Annotation Methods for Creation of a Streamlined, Comprehensive Pipeline." *Genome Biology* 20 (1): 275.

Bibliography

- Pabinger, Stephan, Karina Ernst, Walter Pulverer, Rainer Kallmeyer, Ana M. Valdes, Sarah Metrustry, Denis Katic, et al. 2016. "Analysis and Visualization Tool for Targeted Amplicon Bisulfite Sequencing on Ion Torrent Sequencers." *PloS One* 11 (7): e0160227.
- Pan, Hong, Joanna D. Holbrook, Neerja Karnani, and Chee Keong Kwoh. 2016. "Gene, Environment and Methylation (GEM): A Tool Suite to Efficiently Navigate Large Scale Epigenome Wide Association Studies and Integrate Genotype and Interaction between Genotype and Environment." *BMC Bioinformatics* 17 (August): 299.
- Papareddy, Ranjith K., Katalin Páldi, Anna D. Smolka, Patrick Hüther, Claude Becker, and Michael D. Nodine. 2021. "Repression of CHROMOMETHYLASE 3 Prevents Epigenetic Collateral Damage in Arabidopsis." *eLife* 10 (July). <https://doi.org/10.7554/eLife.69396>.
- Patra Bhattacharya, Deblina, Sebastian Canzler, Stephanie Kehr, Jana Hertel, Ivo Grosse, and Peter F. Stadler. 2016. "Phylogenetic Distribution of Plant snoRNA Families." *BMC Genomics* 17 (1): 969.
- Paul, Dirk S., and Stephan Beck. 2014. "Advances in Epigenome-Wide Association Studies for Common Diseases." *Trends in Molecular Medicine* 20 (10): 541–43.
- Pedersen, Brent S., Kenneth Eyring, Subhajyoti De, Ivana V. Yang, and David A. Schwartz. 2014. "Fast and Accurate Alignment of Long Bisulfite-Seq Reads." *arXiv [q-bio.GN]*. *arXiv*. <http://arxiv.org/abs/1401.1129>.
- Pe'er, Itsik, Roman Yelensky, David Altshuler, and Mark J. Daly. 2008. "Estimation of the Multiple Testing Burden for Genomewide Association Studies of Nearly All Common Variants." *Genetic Epidemiology* 32 (4): 381–85.
- Pérez, Guillermo, Bernard Slippers, Michael J. Wingfield, Brenda D. Wingfield, Angus J. Carnegie, and Treena I. Burgess. 2012. "Cryptic Species, Native Populations and Biological Invasions by a Eucalypt Forest Pathogen." *Molecular Ecology* 21 (18): 4452–71.
- Phillippy, Adam M., Michael C. Schatz, and Mihai Pop. 2008. "Genome Assembly Forensics: Finding the Elusive Mis-Assembly." *Genome Biology* 9 (3): R55.
- Phippen, Winthrop B., and Mary E. Phippen. 2012. "Soybean Seed Yield and Quality as a Response to Field Pennycress Residue." *Crop Science* 52 (6): 2767–73.
- Pignatta, Daniela, Katherine Novitzky, P. R. V. Satyaki, and Mary Gehring. 2018. "A Variably Imprinted Epiallele Impacts Seed Development." *PLoS Genetics* 14 (11): e1007469.

- Pikaard, Craig S., and Ortrun Mittelsten Scheid. 2014. "Epigenetic Regulation in Plants." *Cold Spring Harbor Perspectives in Biology* 6 (12): a019315.
- Platt, Alexander, Paul F. Gugger, Matteo Pellegrini, and Victoria L. Sork. 2015. "Genome-Wide Signature of Local Adaptation Linked to Variable CpG Methylation in Oak Populations." *Molecular Ecology* 24 (15): 3823–30.
- Poplin, Ryan, Valentin Ruano-Rubio, Mark A. DePristo, Tim J. Fennell, Mauricio O. Carneiro, Geraldine A. Van der Auwera, David E. Kling, et al. 2018. "Scaling Accurate Genetic Variant Discovery to Tens of Thousands of Samples." *bioRxiv*. <https://doi.org/10.1101/201178>.
- Poptsova, Maria S., Irina A. Il'icheva, Dmitry Yu Nechipurenko, Larisa A. Panchenko, Mingian V. Khodikov, Nina Y. Oparina, Robert V. Polozov, Yury D. Nechipurenko, and Sergei L. Grokhovsky. 2014. "Non-Random DNA Fragmentation in next-Generation Sequencing." *Scientific Reports* 4 (March): 4532.
- Prezza, Nicola, Cristian Del Fabbro, Francesco Vezzi, Emanuele De Paoli, and Alberto Policriti. 2012. "ERNE-BS5: Aligning BS-Treated Sequences by Multiple Hits on a 5-Letters Alphabet." In *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine*, 12–19. BCB '12. New York, NY, USA: Association for Computing Machinery.
- Prober, J. M., G. L. Trainor, R. J. Dam, F. W. Hobbs, C. W. Robertson, R. J. Zagursky, A. J. Cocuzza, M. A. Jensen, and K. Baumeister. 1987. "A System for Rapid DNA Sequencing with Fluorescent Chain-Terminating Dideoxynucleotides." *Science* 238 (4825): 336–41.
- Pruitt, Kim D., Tatiana Tatusova, Garth R. Brown, and Donna R. Maglott. 2012. "NCBI Reference Sequences (RefSeq): Current Status, New Features and Genome Annotation Policy." *Nucleic Acids Research* 40 (Database issue): D130–35.
- Quinlan, Aaron R., and Ira M. Hall. 2010. "BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features." *Bioinformatics* 26 (6): 841–42.
- Rahmani, Elinor, Reut Yedidim, Liat Shenhav, Regev Schweiger, Omer Weissbrod, Noah Zaitlen, and Eran Halperin. 2017. "GLINT: A User-Friendly Toolset for the Analysis of High-Throughput DNA-Methylation Array Data." *Bioinformatics* 33 (12): 1870–72.
- Raine, Amanda, Ulrika Liljedahl, and Jessica Nordlund. 2018. "Data Quality of Whole Genome Bisulfite Sequencing on Illumina Platforms." *PloS One* 13 (4): e0195972.

Bibliography

- Raineri, Emanuele, Marc Dabad, and Simon Heath. 2014. "A Note on Exact Differences between Beta Distributions in Genomic (Methylation) Studies." *PloS One* 9 (5): e97349.
- Rakyan, Vardhman K., Thomas A. Down, David J. Balding, and Stephan Beck. 2011. "Epigenome-Wide Association Studies for Common Human Diseases." *Nature Reviews. Genetics* 12 (8): 529–41.
- Ramírez, Fidel, Devon P. Ryan, Björn Grüning, Vivek Bhardwaj, Fabian Kilpert, Andreas S. Richter, Steffen Heyne, Friederike Dündar, and Thomas Manke. 2016. "deepTools2: A next Generation Web Server for Deep-Sequencing Data Analysis." *Nucleic Acids Research* 44 (W1): W160–65.
- Ratel, David, Jean-Luc Ravanat, François Berger, and Didier Wion. 2006. "N6-Methyladenine: The Other Methylated Base of DNA." *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology* 28 (3): 309–15.
- Reinders, Jon, Brande B. H. Wulff, Marie Mirouze, Arturo Marí-Ordóñez, Mélanie Dapp, Wilfried Rozhon, Etienne Bucher, Grégory Theiler, and Jerzy Paszkowski. 2009. "Compromised Stability of DNA Methylation and Transposon Immobilization in Mosaic Arabidopsis Epigenomes." *Genes & Development* 23 (8): 939–50.
- Rhie, Arang, Shane A. McCarthy, Olivier Fedrigo, Joana Damas, Giulio Formenti, Sergey Koren, Marcela Uliano-Silva, et al. 2021. "Towards Complete and Error-Free Genome Assemblies of All Vertebrate Species." *Nature* 592 (7856): 737–46.
- Richards. 2006. "Inherited Epigenetic Variation—revisiting Soft Inheritance." *Nature Reviews. Genetics*. <https://www.nature.com/articles/nrg1834>.
- Richards, Christina L., Conchita Alonso, Claude Becker, Oliver Bossdorf, Etienne Bucher, Maria Colomé-Tatché, Walter Durka, et al. 2017. "Ecological Plant Epigenetics: Evidence from Model and Non-Model Species, and the Way Forward." *Ecology Letters* 20 (12): 1576–90.
- Riddle, Nicole C., and Eric J. Richards. 2002. "The Control of Natural Variation in Cytosine Methylation in Arabidopsis." *Genetics* 162 (1): 355–63.
- Riggs, and Porter. 1996. "Overview of Epigenetic Mechanisms." *Cold Spring Harbor Monograph Series*.

- Rimmer, Andy, Hang Phan, Iain Mathieson, Zamin Iqbal, Stephen R. F. Twigg, WGS500 Consortium, Andrew O. M. Wilkie, Gil McVean, and Gerton Lunter. 2014. “Integrating Mapping-, Assembly- and Haplotype-Based Approaches for Calling Variants in Clinical Sequencing Applications.” *Nature Genetics* 46 (8): 912–18.
- Robinson, Mark D., Davis J. McCarthy, and Gordon K. Smyth. 2010. “edgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data.” *Bioinformatics* 26 (1): 139–40.
- Robinson, Mark D., and Alicia Oshlack. 2010. “A Scaling Normalization Method for Differential Expression Analysis of RNA-Seq Data.” *Genome Biology* 11 (3): R25.
- Ronemus, M. J., M. Galbiati, C. Ticknor, J. Chen, and S. L. Dellaporta. 1996. “Demethylation-Induced Developmental Pleiotropy in Arabidopsis.” *Science* 273 (5275): 654–57.
- Rothberg, Jonathan M., Wolfgang Hinz, Todd M. Rearick, Jonathan Schultz, William Mileski, Mel Davey, John H. Leamon, et al. 2011. “An Integrated Semiconductor Device Enabling Non-Optical Genome Sequencing.” *Nature* 475 (7356): 348–52.
- Sáez-Laguna, Enrique, María-Ángeles Guevara, Luis-Manuel Díaz, David Sánchez-Gómez, Carmen Collada, Ismael Aranda, and María-Teresa Cervera. 2014. “Epigenetic Variability in the Genetically Uniform Forest Tree Species *Pinus Pinea* L.” *PloS One* 9 (8): e103145.
- Sahu, Pranav Pankaj, Garima Pandey, Namisha Sharma, Swati Puranik, Mehanathan Muthamilarasan, and Manoj Prasad. 2013. “Epigenetic Mechanisms of Plant Stress Responses and Adaptation.” *Plant Cell Reports* 32 (8): 1151–59.
- Sanger, F., G. M. Air, B. G. Barrell, N. L. Brown, A. R. Coulson, J. C. Fiddes, C. A. Hutchison, P. M. Slocombe, and M. Smith. 1977. “Nucleotide Sequence of Bacteriophage ϕ X174 DNA.” *Nature* 265 (5596): 687–95.
- Sanger, F., S. Nicklen, and A. R. Coulson. 1977. “DNA Sequencing with Chain-Terminating Inhibitors.” *Proceedings of the National Academy of Sciences of the United States of America* 74 (12): 5463–67.
- Saxena, Rachit K., David Edwards, and Rajeev K. Varshney. 2014. “Structural Variations in Plant Genomes.” *Briefings in Functional Genomics* 13 (4): 296–307.

Bibliography

- Saze, Hidetoshi, Ortrun Mittelsten Scheid, and Jerzy Paszkowski. 2003. "Maintenance of CpG Methylation Is Essential for Epigenetic Inheritance during Plant Gametogenesis." *Nature Genetics* 34 (1): 65–69.
- Schatz, Michael C., Jan Witkowski, and W. Richard McCombie. 2012. "Current Challenges in de Novo Plant Genome Sequencing and Assembly." *Genome Biology* 13 (4): 243.
- Scheben, Armin, Yuxuan Yuan, and David Edwards. 2016. "Advances in Genomics for Adapting Crops to Climate Change." *Current Plant Biology* 6 (October): 2–10.
- Schild, Drew R., Matthew R. Walsh, Daren C. Card, Audra L. Andrew, Richard H. Adams, and Todd A. Castoe. 2016. "Epi RAD Seq: Scalable Analysis of Genomewide Patterns of Methylation Using Next-generation Sequencing." *Methods in Ecology and Evolution / British Ecological Society* 7 (1): 60–69.
- Schmitz, Robert J., Matthew D. Schultz, Mathew G. Lewsey, Ronan C. O'Malley, Mark A. Urich, Ondrej Libiger, Nicholas J. Schork, and Joseph R. Ecker. 2011. "Transgenerational Epigenetic Instability Is a Source of Novel Methylation Variants." *Science* 334 (6054): 369–73.
- Schmitz, Robert J., Matthew D. Schultz, Mark A. Urich, Joseph R. Nery, Mattia Pelizzola, Ondrej Libiger, Andrew Alix, et al. 2013. "Patterns of Population Epigenomic Diversity." *Nature* 495 (7440): 193–98.
- Schranz, M. Eric, Martin A. Lysak, and Thomas Mitchell-Olds. 2006. "The ABC's of Comparative Genomics in the Brassicaceae: Building Blocks of Crucifer Genomes." *Trends in Plant Science* 11 (11): 535–42.
- Schulthess, Albert W., Jochen C. Reif, Jie Ling, Jörg Plieske, Sonja Kollers, Erhard Ebmeyer, Viktor Korzun, et al. 2017. "The Roles of Pleiotropy and Close Linkage as Revealed by Association Mapping of Yield and Correlated Traits of Wheat (*Triticum Aestivum* L.)." *Journal of Experimental Botany* 68 (15): 4089–4101.
- Schultz, Matthew D., Yupeng He, John W. Whitaker, Manoj Hariharan, Eran A. Mukamel, Danny Leung, Nisha Rajagopal, et al. 2016. "Corrigendum: Human Body Epigenome Maps Reveal Noncanonical DNA Methylation Variation." *Nature* 530 (7589): 242.
- Schultz, Matthew D., Robert J. Schmitz, and Joseph R. Ecker. 2012. "'Leveling' the Playing Field for Analyses of Single-Base Resolution DNA Methylomes." *Trends in Genetics: TIG* 28 (12): 583–85.

- Secco, David, Chuang Wang, Huixia Shou, Matthew D. Schultz, Serge Chiarenza, Laurent Nussaume, Joseph R. Ecker, James Whelan, and Ryan Lister. 2015. "Stress Induced Gene Expression Drives Transient DNA Methylation Changes at Adjacent Repetitive Elements." *eLife* 4 (July). <https://doi.org/10.7554/eLife.09343>.
- Sedbrook, John C., Winthrop B. Phippen, and M. David Marks. 2014. "New Approaches to Facilitate Rapid Domestication of a Wild Plant to an Oilseed Crop: Example Pennycress (*Thlaspi Arvense* L.)." *Plant Science: An International Journal of Experimental Plant Biology* 227 (October): 122–32.
- Seper, Bernice, Krista van den Heuvel, Melanie Lindner, Heidi Viitaniemi, Arild Husby, and Kees van Oers. 2019. "Avian Ecological Epigenetics: Pitfalls and Promises." *Journal of Ornithology / DO-G* 160 (4): 1183–1203.
- Shabalin, Andrey A. 2012. "Matrix eQTL: Ultra Fast eQTL Analysis via Large Matrix Operations." *Bioinformatics* 28 (10): 1353–58.
- Shapiro, Robert, Robert E. Servis, and Marvin Welcher. 1970. "Reactions of Uracil and Cytosine Derivatives with Sodium Bisulfite." *Journal of the American Chemical Society* 92 (2): 422–24.
- Sheldon, C. C., J. E. Burn, P. P. Perez, J. Metzger, J. A. Edwards, W. J. Peacock, and E. S. Dennis. 1999. "The FLF MADS Box Gene: A Repressor of Flowering in *Arabidopsis* Regulated by Vernalization and Methylation." *The Plant Cell* 11 (3): 445–58.
- Sheldon, C. C., D. T. Rouse, E. J. Finnegan, W. J. Peacock, and E. S. Dennis. 2000. "The Molecular Basis of Vernalization: The Central Role of FLOWERING LOCUS C (FLC)." *Proceedings of the National Academy of Sciences of the United States of America* 97 (7): 3753–58.
- Shendure, Jay, Gregory J. Porreca, Nikos B. Reppas, Xiaoxia Lin, John P. McCutcheon, Abraham M. Rosenbaum, Michael D. Wang, Kun Zhang, Robi D. Mitra, and George M. Church. 2005. "Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome." *Science* 309 (5741): 1728–32.
- Shumate, Alaina, and Steven L. Salzberg. 2020. "Liftoff: Accurate Mapping of Gene Annotations." *Bioinformatics*, December. <https://doi.org/10.1093/bioinformatics/btaa1016>.
- Sigman, Meredith J., and R. Keith Slotkin. 2016. "The First Rule of Plant Transposable Element Silencing: Location, Location, Location." *The Plant Cell* 28 (2): 304–13.

Bibliography

- Simão, Felipe A., Robert M. Waterhouse, Panagiotis Ioannidis, Evgenia V. Kriventseva, and Evgeny M. Zdobnov. 2015. "BUSCO: Assessing Genome Assembly and Annotation Completeness with Single-Copy Orthologs." *Bioinformatics* 31 (19): 3210–12.
- Šimková, H. 1998. "Methylation of Mitochondrial DNA in Carrot (*Daucus Carota* L.)." *Plant Cell Reports* 17 (3): 220–24.
- Slotkin, R. Keith. 2016. "Plant Epigenetics: From Genotype to Phenotype and Back Again." *Genome Biology* 17 (March): 57.
- Smit, Arian F. A. 2004. "Repeat-Masker Open-3.0." <http://www.repeatmasker.org>.
<https://ci.nii.ac.jp/naid/10029514778/>.
- Smit, Arian F. A., and Robert Hubley. 2008. "RepeatModeler Open-1.0."
- Smith, L. M., J. Z. Sanders, R. J. Kaiser, P. Hughes, C. Dodd, C. R. Connell, C. Heiner, S. B. Kent, and L. E. Hood. 1986. "Fluorescence Detection in Automated DNA Sequence Analysis." *Nature* 321 (6071): 674–79.
- Smith, T. F., and M. S. Waterman. 1981. "Identification of Common Molecular Subsequences." *Journal of Molecular Biology* 147 (1): 195–97.
- Sonnhammer, E. L., and R. Durbin. 1995. "A Dot-Matrix Program with Dynamic Threshold Control Suited for Genomic DNA and Protein Sequence Analysis." *Gene* 167 (1-2): GC1–10.
- Srivastava, Akanksha, Yuliya V. Karpievitch, Steven R. Eichten, Justin O. Borevitz, and Ryan Lister. 2019. "HOME: A Histogram Based Machine Learning Approach for Effective Identification of Differentially Methylated Regions." *BMC Bioinformatics* 20 (1): 253.
- Stam, Piet. 1993. "Construction of Integrated Genetic Linkage Maps by Means of a New Computer Package: Join Map." *The Plant Journal: For Cell and Molecular Biology* 3 (5): 739–44.
- Stanke, Mario, Oliver Keller, Irfan Gunduz, Alec Hayes, Stephan Waack, and Burkhard Morgenstern. 2006. "AUGUSTUS: Ab Initio Prediction of Alternative Transcripts." *Nucleic Acids Research* 34 (Web Server issue): W435–39.
- Stanke, Mario, and Stephan Waack. 2003. "Gene Prediction with a Hidden Markov Model and a New Intron Submodel." *Bioinformatics* 19 Suppl 2 (October): ii215–25.

- Stroud, Hume, Bo Ding, Stacey A. Simon, Suhua Feng, Maria Bellizzi, Matteo Pellegrini, Guo-Liang Wang, Blake C. Meyers, and Steven E. Jacobsen. 2013. "Plants Regenerated from Tissue Culture Contain Stable Epigenome Changes in Rice." *eLife* 2 (March): e00354.
- Stroud, Hume, Truman Do, Jiamu Du, Xuehua Zhong, Suhua Feng, Lianna Johnson, Dinshaw J. Patel, and Steven E. Jacobsen. 2014. "Non-CG Methylation Patterns Shape the Epigenetic Landscape in Arabidopsis." *Nature Structural & Molecular Biology* 21 (1): 64–72.
- Sultana, Tania, Alessia Zamborlini, Gael Cristofari, and Pascale Lesage. 2017. "Integration Site Selection by Retroviruses and Transposable Elements in Eukaryotes." *Nature Reviews. Genetics* 18 (5): 292–308.
- Suzuki, Liao, Wos, Johnston, DeGrazia, Ishii, Bloom, Zody, Germer, and Grealley. 2018. "Whole-Genome Bisulfite Sequencing with Improved Accuracy and Cost." *Genome Research/ a Journal of Science and Its Applications*. <https://genome.cshlp.org/content/28/9/1364.short>.
- Suzuki, Miho M., and Adrian Bird. 2008. "DNA Methylation Landscapes: Provocative Insights from Epigenomics." *Nature Reviews. Genetics* 9 (6): 465–76.
- Tang, Haibao, John E. Bowers, Xiyin Wang, Ray Ming, Maqsoodul Alam, and Andrew H. Paterson. 2008. "Synteny and Collinearity in Plant Genomes." *Science* 320 (5875): 486–88.
- Tang, Haibao, Xingtang Zhang, Chenyong Miao, Jisen Zhang, Ray Ming, James C. Schnable, Patrick S. Schnable, Eric Lyons, and Jianguo Lu. 2015. "ALLMAPS: Robust Scaffold Ordering Based on Multiple Maps." *Genome Biology* 16 (January): 3.
- Tedeschi, Francesca, Paride Rizzo, Bui Thi Mai Huong, Andreas Czihal, Twan Rutten, Lothar Altschmied, Sarah Scharfenberg, et al. 2019. "EFFECTOR OF TRANSCRIPTION Factors Are Novel Plant-Specific Regulators Associated with Genomic DNA Methylation in Arabidopsis." *The New Phytologist* 221 (1): 261–78.
- Thomas, Jason B., Marshall E. Hampton, Kevin M. Dorn, M. David Marks, and Clay J. Carter. 2017. "The Pennycress (*Thlaspi Arvense* L.) Nectary: Structural and Transcriptomic Characterization." *BMC Plant Biology* 17 (1): 201.
- Thornsberry, J. M., M. M. Goodman, J. Doebley, S. Kresovich, D. Nielsen, and E. S. Buckler. 2001. "Dwarf8 Polymorphisms Associate with Variation in Flowering Time." *Nature Genetics* 28 (3): 286–89.

Bibliography

- Tibbs Cortes, Laura, Zhiwu Zhang, and Jianming Yu. 2021. "Status and Prospects of Genome-Wide Association Studies in Plants." *The Plant Genome* 14 (1): e20077.
- Tomato Genome Consortium. 2012. "The Tomato Genome Sequence Provides Insights into Fleshy Fruit Evolution." *Nature* 485 (7400): 635–41.
- Tran, Hong, Jacob Porter, Ming-An Sun, Hehuang Xie, and Liqing Zhang. 2014. "Objective and Comprehensive Evaluation of Bisulfite Short Read Mapping Tools." *Advances in Bioinformatics* 2014 (April): 472045.
- Treangen, Todd J., and Steven L. Salzberg. 2011. "Repetitive DNA and next-Generation Sequencing: Computational Challenges and Solutions." *Nature Reviews. Genetics* 13 (1): 36–46.
- Trucchi, Emiliano, Anna B. Mazzarella, Gregor D. Gilfillan, Maria T. Lorenzo, Peter Schönswetter, and Ovidiu Paun. 2016. "BsRADseq: Screening DNA Methylation in Natural Populations of Non-Model Species." *Molecular Ecology* 25 (8): 1697–1713.
- Tsuji, Junko, and Zhiping Weng. 2016. "Evaluation of Preprocessing, Mapping and Postprocessing Algorithms for Analyzing Whole Genome Bisulfite Sequencing Data." *Briefings in Bioinformatics* 17 (6): 938–52.
- Tsukahara, Sayuri, Akie Kobayashi, Akira Kawabe, Olivier Mathieu, Asuka Miura, and Tetsuji Kakutani. 2009. "Bursts of Retrotransposition Reproduced in Arabidopsis." *Nature* 461 (7262): 423–26.
- Tuskan, G. A., S. Difazio, S. Jansson, J. Bohlmann, I. Grigoriev, U. Hellsten, N. Putnam, et al. 2006. "The Genome of Black Cottonwood, *Populus Trichocarpa* (Torr. & Gray)." *Science* 313 (5793): 1596–1604.
- Uemura, Sotaro, Colin Echeverría Aitken, Jonas Korlach, Benjamin A. Flusberg, Stephen W. Turner, and Joseph D. Puglisi. 2010. "Real-Time tRNA Transit on Single Translating Ribosomes at Codon Resolution." *Nature* 464 (7291): 1012–17.
- Van Der Graaf, and Wardenaar. 2015. "Rate, Spectrum, and Evolutionary Dynamics of Spontaneous Epimutations." *Proceedings of the Estonian Academy of Sciences. Biology, Ecology = Eesti Teaduste Akadeemia Toimetised. Bioloogia, Okoloogia.* <https://www.pnas.org/content/112/21/6676.short>.

- Vanyushin, B. F., and M. D. Kirnos. 1988. "DNA Methylation in Plants." *Gene* 74 (1): 117–21.
- Vaser, Robert, Ivan Sović, Niranjana Nagarajan, and Mile Šikić. 2017. "Fast and Accurate de Novo Genome Assembly from Long Uncorrected Reads." *Genome Research* 27 (5): 737–46.
- Vaughn, Matthew W., Milos Tanurdzić, Zachary Lippman, Hongmei Jiang, Robert Carrasquillo, Pablo D. Rabinowicz, Neilay Dedhia, et al. 2007. "Epigenetic Natural Variation in *Arabidopsis Thaliana*." *PLoS Biology* 5 (7): e174.
- Venney, Clare J., Mattias L. Johansson, and Daniel D. Heath. 2016. "Inbreeding Effects on Gene-Specific DNA Methylation among Tissues of Chinook Salmon." *Molecular Ecology* 25 (18): 4521–33.
- Visscher, Peter M., Naomi R. Wray, Qian Zhang, Pamela Sklar, Mark I. McCarthy, Matthew A. Brown, and Jian Yang. 2017. "10 Years of GWAS Discovery: Biology, Function, and Translation." *American Journal of Human Genetics* 101 (1): 5–22.
- Vitte, Clémentine, Margaux-Alison Fustier, Karine Alix, and Maud I. Tenailon. 2014. "The Bright Side of Transposons in Crop Evolution." *Briefings in Functional Genomics* 13 (4): 276–95.
- Voinnet, Olivier. 2009. "Origin, Biogenesis, and Activity of Plant microRNAs." *Cell* 136 (4): 669–87.
- Wagner, I., and I. Capesius. 1981. "Determination of 5-Methylcytosine from Plant DNA by High-Performance Liquid Chromatography." *Biochimica et Biophysica Acta* 654 (1): 52–56.
- Walker, Bruce J., Thomas Abeel, Terrance Shea, Margaret Priest, Amr Abouelliel, Sharadha Sakthikumar, Christina A. Cuomo, et al. 2014. "Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement." *PloS One* 9 (11): e112963.
- Wang, Xiaowu, Hanzhong Wang, Jun Wang, Rifei Sun, Jian Wu, Shengyi Liu, Yinqi Bai, et al. 2011. "The Genome of the Mesopolyploid Crop Species *Brassica Rapa*." *Nature Genetics* 43 (10): 1035–39.
- Warnes, Bolker, Bonebakker, and Gentleman. 2009. "Gplots: Various R Programming Tools for Plotting Data." R Package Version.
- Warwick, S. I., A. Francis, and D. J. Susko. 2002. "The Biology of Canadian Weeds. 9. *Thlaspi Arvense* L. (updated)." *Canadian Journal of Plant Science. Revue Canadienne de Phytotechnie* 82 (4): 803–23.

Bibliography

- Wassenegger, M., S. Heimes, L. Riedel, and H. L. Sanger. 1994. "RNA-Directed de Novo Methylation of Genomic Sequences in Plants." *Cell* 76 (3): 567–76.
- Weese, David, Manuel Holtgrewe, and Knut Reinert. 2012. "RazerS 3: Faster, Fully Sensitive Read Mapping." *Bioinformatics* 28 (20): 2592–99.
- Wee, Yongkiat, Salma Begum Bhyan, Yining Liu, Jiachun Lu, Xiaoyan Li, and Min Zhao. 2019. "The Bioinformatics Tools for the Genome Assembly and Analysis Based on Third-Generation Sequencing." *Briefings in Functional Genomics* 18 (1): 1–12.
- Wendel, Jonathan F. 2015. "The Wondrous Cycles of Polyploidy in Plants." *American Journal of Botany* 102 (11): 1753–56.
- Wendel, Jonathan F., Scott A. Jackson, Blake C. Meyers, and Rod A. Wing. 2016. "Evolution of Plant Genome Architecture." *Genome Biology* 17 (March): 37.
- Wendte, Jered M., Yinwen Zhang, Lexiang Ji, Xiuling Shi, Rashmi R. Hazarika, Yadollah Shahryary, Frank Johannes, and Robert J. Schmitz. 2019. "Epimutations Are Associated with CHROMOMETHYLASE 3-Induced de Novo DNA Methylation." *eLife* 8 (July). <https://doi.org/10.7554/eLife.47891>.
- Werner, Olaf, ngela S. Prudencio, Elena de la Cruz-Martnez, Marta Nieto-Lugilde, Pedro Martnez-Gmez, and Rosa M. Ros. 2020. "A Cost Reduced Variant of Epi-Genotyping by Sequencing for Studying DNA Methylation in Non-Model Organisms." *Frontiers in Plant Science* 11 (May): 694.
- Weyers, Sharon L., Russ W. Gesch, Frank Forcella, Carrie A. Eberle, Matthew D. Thom, Heather L. Matthees, Matthew Ott, Gary W. Feyereisen, and Jeffrey S. Strock. 2021. "Surface Runoff and Nutrient Dynamics in Cover Crop-Soybean Systems in the Upper Midwest." *Journal of Environmental Quality* 50 (1): 158–71.
- Weyers, Sharon, Matt Thom, Frank Forcella, Carrie Eberle, Heather Matthees, Russ Gesch, Matthew Ott, Gary Feyereisen, Jeffery Strock, and Don Wyse. 2019. "Reduced Potential for Nitrogen Loss in Cover Crop-Soybean Relay Systems in a Cold Climate." *Journal of Environmental Quality* 48 (3): 660–69.
- Wibowo, Anjar, Claude Becker, Julius Durr, Jonathan Price, Stijn Spaepen, Sally Hilton, Hadi Putra, et al. 2018. "Partial Maintenance of Organ-Specific Epigenetic Marks during Plant Asexual

Reproduction Leads to Heritable Phenotypic Variation.” *Proceedings of the National Academy of Sciences of the United States of America* 115 (39): E9145–52.

Wibowo, Anjar, Claude Becker, Gianpiero Marconi, Julius Durr, Jonathan Price, Jorg Hagmann, Ranjith Papareddy, et al. 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 (May). <https://doi.org/10.7554/eLife.13546>.

Wickham, Hadley. 2011. “Ggplot2.” *Wiley Interdisciplinary Reviews. Computational Statistics* 3 (2): 180–85.

Workman, Rachael E., Alexander M. Myrka, G. William Wong, Elizabeth Tseng, Kenneth C. Welch Jr, and Winston Timp. 2018. “Single-Molecule, Full-Length Transcript Sequencing Provides Insight into the Extreme Metabolism of the Ruby-Throated Hummingbird *Archilochus Colubris*.” *GigaScience* 7 (3): 1–12.

Workman, Rachael, Renee Fedak, Duncan Kilburn, Stephanie Hao, Kelvin Liu, and Winston Timp. 2019. “High Molecular Weight DNA Extraction from Recalcitrant Plant Species for Third Generation Sequencing v1.” <https://doi.org/10.17504/protocols.io.4vbgw2n>.

Wreczycka, Katarzyna, Alexander Goshchan, Dilmurat Yusuf, Björn Grüning, Yassen Assenov, and Altuna Akalin. 2017. “Strategies for Analyzing Bisulfite Sequencing Data.” *Journal of Biotechnology* 261 (November): 105–15.

Wu, Thomas D., and Serban Nacu. 2010. “Fast and SNP-Tolerant Detection of Complex Variants and Splicing in Short Reads.” *Bioinformatics* 26 (7): 873–81.

Xie, H. J., H. Li, D. Liu, W. M. Dai, J. Y. He, S. Lin, H. Duan, et al. 2015. “ICE1 Demethylation Drives the Range Expansion of a Plant Invader through Cold Tolerance Divergence.” *Molecular Ecology* 24 (4): 835–50.

Xie, Shangqian, Amy Wing-Sze Leung, Zhenxian Zheng, Dake Zhang, Chuanle Xiao, Ruibang Luo, Ming Luo, and Shoudong Zhang. 2021. “Applications and Potentials of Nanopore Sequencing in the (epi)genome and (epi)transcriptome Era.” *The Innovation Journal: The Public Sector Innovation Journal* 2 (4): 100153.

Xi, Yuanxin, and Wei Li. 2009. “BSMAP: Whole Genome Bisulfite Sequence MAPping Program.” *BMC Bioinformatics* 10 (July): 232.

Bibliography

- Xu, Jing, Linna Zhao, Di Liu, Simeng Hu, Xiuling Song, Jin Li, Hongchao Lv, et al. 2018. "EWAS: Epigenome-Wide Association Study Software 2.0." *Bioinformatics* 34 (15): 2657–58.
- Xu, Yong-Chao, Xiao-Min Niu, Xin-Xin Li, Wenrong He, Jia-Fu Chen, Yu-Pan Zou, Qiong Wu, Yong E. Zhang, Wolfgang Busch, and Ya-Long Guo. 2019. "Adaptation and Phenotypic Diversification in Arabidopsis through Loss-of-Function Mutations in Protein-Coding Genes." *The Plant Cell* 31 (5): 1012–25.
- Yanai, Itai, Hila Benjamin, Michael Shmoish, Vered Chalifa-Caspi, Maxim Shklar, Ron Ophir, Arren Bar-Even, et al. 2005. "Genome-Wide Midrange Transcription Profiles Reveal Expression Level Relationships in Human Tissue Specification." *Bioinformatics* 21 (5): 650–59.
- Yang, Ruolin, David E. Jarvis, Hao Chen, Mark A. Beilstein, Jane Grimwood, Jerry Jenkins, Shengqiang Shu, et al. 2013. "The Reference Genome of the Halophytic Plant *Eutrema Salsugineum*." *Frontiers in Plant Science* 4 (March): 46.
- Zemach, Assaf, M. Yvonne Kim, Ping-Hung Hsieh, Devin Coleman-Derr, Leor Eshed-Williams, Ka Thao, Stacey L. Harmer, and Daniel Zilberman. 2013. "The Arabidopsis Nucleosome Remodeler DDM1 Allows DNA Methyltransferases to Access H1-Containing Heterochromatin." *Cell* 153 (1): 193–205.
- Zemach, Assaf, Ivy E. McDaniel, Pedro Silva, and Daniel Zilberman. 2010. "Genome-Wide Evolutionary Analysis of Eukaryotic DNA Methylation." *Science* 328 (5980): 916–19.
- Zhang, Huiming, Zhaobo Lang, and Jian-Kang Zhu. 2018. "Dynamics and Function of DNA Methylation in Plants." *Nature Reviews. Molecular Cell Biology* 19 (8): 489–506.
- Zhang, Mei, Shaojun Xie, Xiaomei Dong, Xin Zhao, Biao Zeng, Jian Chen, Hui Li, et al. 2014. "Genome-Wide High Resolution Parental-Specific DNA and Histone Methylation Maps Uncover Patterns of Imprinting Regulation in Maize." *Genome Research* 24 (1): 167–76.
- Zhang, Qingzhu, Dong Wang, Zhaobo Lang, Li He, Lan Yang, Liang Zeng, Yanqiang Li, et al. 2016. "Methylation Interactions in Arabidopsis Hybrids Require RNA-Directed DNA Methylation and Are Influenced by Genetic Variation." *Proceedings of the National Academy of Sciences of the United States of America* 113 (29): E4248–56.
- Zhang, Shi-Jian, Lei Liu, Ruolin Yang, and Xiangfeng Wang. 2020. "Genome Size Evolution Mediated by Gypsy Retrotransposons in Brassicaceae." *Genomics, Proteomics & Bioinformatics* 18 (3): 321–32.

- Zhang, Xiaoyu, Junshi Yazaki, Ambika Sundaresan, Shawn Cokus, Simon W-L Chan, Huaming Chen, Ian R. Henderson, et al. 2006. "Genome-Wide High-Resolution Mapping and Functional Analysis of DNA Methylation in Arabidopsis." *Cell* 126 (6): 1189–1201.
- Zhang, Xiongfei, Joel Goodsell, and Robert B. Norgren Jr. 2012. "Limitations of the Rhesus Macaque Draft Genome Assembly and Annotation." *BMC Genomics* 13 (May): 206.
- Zhang, Yinwen, Hosung Jang, Rui Xiao, Ioanna Kakoulidou, Robert S. Piecyk, Frank Johannes, and Robert J. Schmitz. 2021. "Heterochromatin Is a Quantitative Trait Associated with Spontaneous Epiallele Formation." *Nature Communications* 12 (1): 6958.
- Zhong, Silin, Zhangjun Fei, Yun-Ru Chen, Yi Zheng, Mingyun Huang, Julia Vrebalov, Ryan McQuinn, et al. 2013. "Single-Base Resolution Methyomes of Tomato Fruit Development Reveal Epigenome Modifications Associated with Ripening." *Nature Biotechnology* 31 (2): 154–59.
- Zhou, Qiangwei, Ze Wang, Jing Li, Wing-Kin Sung, and Guoliang Li. 2020. "MethHaplo: Combining Allele-Specific DNA Methylation and SNPs for Haplotype Region Identification." *BMC Bioinformatics* 21 (1): 451.
- Zilberman, Daniel, Mary Gehring, Robert K. Tran, Tracy Ballinger, and Steven Henikoff. 2007. "Genome-Wide Analysis of Arabidopsis Thaliana DNA Methylation Uncovers an Interdependence between Methylation and Transcription." *Nature Genetics* 39 (1): 61–69.
- Ziller, Michael J., Hongcang Gu, Fabian Müller, Julie Donaghey, Linus T-Y Tsai, Oliver Kohlbacher, Philip L. De Jager, et al. 2013. "Charting a Dynamic DNA Methylation Landscape of the Human Genome." *Nature* 500 (7463): 477–81.
- Ziller, Michael J., Kasper D. Hansen, Alexander Meissner, and Martin J. Aryee. 2014. "Coverage Recommendations for Methylation Analysis by Whole-Genome Bisulfite Sequencing." *Nature Methods* 12 (3): 230–32.
- Zook, Justin M., Brad Chapman, Jason Wang, David Mittelman, Oliver Hofmann, Winston Hide, and Marc Salit. 2014. "Integrating Human Sequence Data Sets Provides a Resource of Benchmark SNP and Indel Genotype Calls." *Nature Biotechnology* 32 (3): 246–51.
- Zou, Cheng, Melissa D. Lehti-Shiu, Françoise Thibaud-Nissen, Tanmay Prakash, C. Robin Buell, and Shin-Han Shiu. 2009. "Evolutionary and Expression Signatures of Pseudogenes in Arabidopsis and Rice." *Plant Physiology* 151 (1): 3–15.

Curriculum Scientiae

Education

since 04/2018

University of Leipzig, Germany

- PhD student in group of Prof. Peter F. Stadler, Chair of Bioinformatics
- Thesis: *Advancing the analysis of bisulfite sequencing data in its application to ecological plant epigenetics*

09/2015 – 02/2018

University of Lund, Sweden

- Master of Science, Biology (double major)
- Master of Bioinformatics (double major)
- Thesis: *Elucidating the molecular mechanism behind the behavioural manipulation of 'zombie'-flies by the entomopathogenic fungus, Entomophthora muscae*

09/2008 – 06/2011

University of Portsmouth, UK

- Bachelor of Science, Biology/Biological Sciences
- Thesis: *Structure and function of a restriction-modification controller enzyme*

Work Experience

since 04/2018

ecSeq Bioinformatics GmbH, Leipzig, Germany

- EU-Forscher and Lecturer
- Lecturing: *Workshops*

06/2014 – 08/2015

Exosect Limited, Southampton, UK

- Senior Research Scientist
- Performing GLP/GEP experiments for regulatory submission of commercial products
- Project: *Harnessing natural fungi to control insect and mite pests in grain storage*

08/2012 – 05/2014

- Research Scientist
- Performing non-regulatory experiments for research on commercial products

12/2011 – 07/2012

- Research Technician

10/2011 – 12/2011

University of Portsmouth, UK

- Pharmaceutical Sciences Technician

IT-Knowledge

Operating Systems:

UNIX, Mac, Linux, Windows

Programming:

Python, R, Java, Groovy, Nextflow, awk, PHP, BASH

Markup Languages:

LATEX, HTML, Markdown

Database Systems:

MySQL

Languages

English:

Native speaker

Publication Record

Peer-reviewed articles (9)

[Nunn A](#), Otto C, Fasold M, Stadler PF, Langenberger D (2022).

Manipulating base quality scores enables variant calling from bisulfite sequencing alignments using conventional Bayesian approaches. *BMC Genomics*. June; 23, p.477, doi:10.1186/s12864-022-08691-6

Galanti D, Ramos-Cruz D, [Nunn A](#), Rodríguez-Arévalo I, Scheepens JF, Becker C, Bossdorf O (2022).

Genetic and environmental drivers of large-scale epigenetic variation in *Thlaspi arvense*. *PLoS Genetics* [In review]. doi:10.1101/2022.03.16.484610

Gawehns F, Postuma M, van Antro M, [Nunn A](#), Sepers B, Fatma S, van Gurp TP, Wagemaker NCAM, Mateman C, Milanovic-Ivanovic S, Grosse I, van Oers K, Vergeer P, Verhoeven KJF (2022).

epiGBS2: Improvements and evaluation of highly multiplexed, epiGBS-based reduced representation bisulfite sequencing. *Molecular Ecology Resources*. Feb, 22:(5), p.2087-2104, doi:10.1111/1755-0998.13597

Hüther P, Hagmann J, [Nunn A](#), Kakoulidou I, Pisupati R, Langenberger D, Weigel D, Johannes F, Schultheiss SJ, Becker C (2022).

MethylScore: a pipeline for accurate and context-aware identification of differentially methylated regions from population-scale plant WGBS data. *Quantitative Plant Biology* [In review]. doi:10.1101/2022.01.06.475031

[Nunn A](#), Rodríguez-Arévalo I, Tandukar Z, Frels K, Contreras-Garrido A, Carbonell-Bejerano P, Zhang P, Ramos-Cruz D, Jandrasits K, Lanz C, Brusa A, Mirouze M, Dorn K, Jarvis B, Sedbrook J, Wyse DL, Otto C, Langenberger D, Weigel D, Marks MD, Anderson JA, Becker C, Chopra R (2022).

Chromosome-level *Thlaspi arvense* genome provides new tools for translational research and for a newly domesticated cash cover crop of the cooler climates. *Plant Biotechnology Journal*. Jan; 20:(5), p.944-963, doi:10.1111/pbi.13775

[Nunn A](#), Can SN, Otto C, Fasold M, Díez Rodríguez B, Fernandez-Pozo N, Rensing SA, Stadler PF, Langenberger D (2021).

EpiDiverse Toolkit: a pipeline suite for the analysis of ecological plant epigenetics. *NAR Genomics and Bioinformatics*. Dec; lqab106, doi:10.1093/nargab/lqab106

Can SN, [Nunn A](#), Galanti D, Langenberger D, Becker C, Volmer K, Heer K, Opgenoorth L, Fernandez-Pozo N, Rensing SA (2021).

The EpiDiverse Plant Epigenome-Wide Association Studies (EWAS) Pipeline. *Epigenomes*. May; 5:(2) p.12, doi:10.3390/epigenomes5020012

Nunn A, Otto C, Stadler PF, Langenberger D (2021).

Comprehensive benchmarking of software for mapping whole genome bisulfite data: from read alignment to DNA methylation analysis. *Briefings in Bioinformatics*. Feb; bbab021, doi:10.1093/bib/bbab021

Storm C, Scoates F, Nunn A, Potin O, Dillon A (2016).

Improving efficacy of *Beauveria bassiana* against stored grain beetles with a synergistic co-formulant. *Insects*. August; 7:(3), p.42, doi: 10.3390/insects7030042

Miscellaneous (1)

Nunn A (2021).

Bisulfite sequencing methods. In C. Lampei, K. Heer & L. Opgenoorth (Eds.), *Introduction to Ecological Plant Epigenetics*. https://epidiverse.gitbook.io/project/-MfxkdBDZggX_vc_sG5l

Collaborations

EpiDiverse European Training Network, Netherlands Institute of Ecology (NIOO-KNAW).
EU Horizon 2020 program under Marie Skłodowska-Curie grant agreement No 764965.

Prof. Dr. Claude Becker, Genetics, Ludwig-Maximilians-University (LMU) Munich, Germany.

Dr. Ratan Chopra, Dept. of Agronomy and Plant Genetics. University of Minnesota, USA.

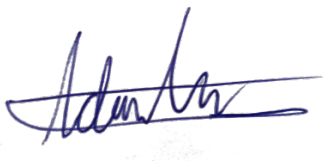
Reviewer Activities

- **Briefings in Bioinformatics**, Oxford University Press
- **BMC Bioinformatics**, BioMed Central
- **Molecular Ecology**, Wiley Online Library

Selbständigkeitserklärung

Hiermit erkläre ich, die vorliegende Dissertation selbständig und ohne unzulässige fremde Hilfe angefertigt zu haben. Ich habe keine anderen als die angeführten Quellen und Hilfsmittel benutzt und sämtliche Textstellen, die wörtlich oder sinngemäß aus veröffentlichten oder unveröffentlichten Schriften entnommen wurden, und alle Angaben, die auf mündlichen Auskünften beruhen, als solche kenntlich gemacht. Ebenfalls sind alle von anderen Personen bereitgestellten Materialien oder Dienstleistungen als solche gekennzeichnet.

Leipzig, den 30. Juni 2022



(Adam Nunn)