Organic or conventional?

Developing a method for food fraud detection based on DNA methylation patterns

Dissertation zur Erlangung des

Doktorgrades der Naturwissenschaften (Dr. rer. nat.)

der

Naturwissenschaftlichen Fakultät III
Agrar- und Ernährungswissenschaften,
Geowissenschaften und Informatik der

MARTIN-LUTHER-UNIVERSITÄT
HALLE-WITTENBERG

vorgelegt von
Herrn M. Sc. Claudius Grehl
Geb. am 24.10.1991 in Leipzig

Erster Gutachter: Prof. Dr. Bruno Glaser

Zweiter Gutachter: Dr. habil. Helmut Bäumlein

Verteidigungstermin: 31.01.2022

I. Table of Contents

II. Abbreviations

5hmC	5-hydroxymethyl-Cytosine
5mC	5-methyl-Cytosine
A	Adenine
bp	base pairs
C	Cytosine
CpG	Cytosine-phosphate-Guanine-dinucleotide
DMC	Differentially Methylated Cytosines
DMR	Differentially Methylated Regions
DNA	Deoxyribonucleic Acid
G	Guanine
GC/MST	Gas-Chromatography/Mass-Spectrometry Technique
GO	Gene Ontology
H	Adenine, Thymine or Cytosine
ICP-AES	Inductively-Coupled Plasma-Atomic-Emission-Spectrometry
LTR	Long, Terminal Repeats
MeDIPSeq	Methylated DNA Immuno-Precipitation Sequencing
MSRE	Methylation-Sensitive Restriction Enzyme digestion
nt	nucleotides
PBAT	Post-Bisulfite Adapter Tagging
PCR	Polymerase Chain Reaction
RdDM	RNA-directed DNA Methylation
RNA	Ribonucleic Acid
rrBS	Reduced Representation Bisulfite Sequencing
T	Thymine
TGS	Transcriptional Gene Silencing
TSS	Transcription Start Site
WGBS	Whole-Genome Bisulfite-Sequencing

III. List of Tables

IV. List of Figures

V. Summary

The topic of product authentication is of importance as food fraud could, up to now, often not be detected due to the imperfection of currently used methods.

The question of DNA methylation-based product authentication in organic and conventional grown potatoes *(Solanum tuberosum, var. Desirée)* using a Post-Bisulfite-Adapter-Tagging Whole-genome Bisulfit-Sequencing (PBAT-WGBS) protocol is addressed in this dissertation. This new approach uses the environment-dependent addition of methyl-groups to cytosines within the DNA to search for differences between samples grown under varying agricultural managing systems. The samples were grown in the DOK-field trial (Agroscope and Research Institute of Organic Agriculture, Frick, Switzerland). They were raised under field conditions, in three field replicated plots and two years using the same variety and the same seed tuber batches for conventional and bio-organic plots. The bio-organic samples underwent copper-applications for fungi pest control and organic fertilization, while the conventional samples have been treated with synthetic fungicides and fertilized with mineral and organic fertilizers.

The thesis is divided into three parts, wherein the first and the second part were published in peer-reviewed, international scientific journals already. For the third part, the publication is planned.

In a peer-reviewed methodological review article, steps and data analysis currently applied in the field of DNA methylation analysis were described. For our experiment, PBAT-WGBS was chosen due to its reliability, single-base resolution and capability to use low amounts of DNA as input.

To analyse the respective datasets from PBAT-WGBS, a reliable pipeline was needed to bioinformatically evaluate the methylation patterns. This has been implemented in the large-scale pipeline analysis tool snakemake. After an intensive quality benchmark of mapping tools, the alignment was performed with Bismarkbwt2. The results of this benchmark study with 8 sequencing read mapping tools and various parameter set combinations were published in a peer-reviewed article, too.

Finally, differently methylated cytosines (DMC) and differentially methylated regions (DMR) in both years examined were found to be related to metabolic, transport and hormone-related genomic pathways. One DMR in one single year was associated with ATOX1. This indicates a detoxification of copper in organic potatoes but has to be addressed in future studies.

The method was exemplary applied in potatoes grown under organic and conventional farming systems. Besides this application, it offers the potential to analytically monitor storage, transport, cultivation and processing conditions of multiple kinds of products and therefore might be of importance in future product authentication.

In summary, the reliability of the method applied was demonstrated.

VI. Zusammenfassung

Analytische Produktauthentifizierung ist von Bedeutung, da Lebensmittelbetrug aufgrund der Unvollkommenheit der derzeit verwendeten Methoden bisher oft nicht erkannt werden kann.

Die DNA-Methylierungs-basierte Produktauthentifizierung von biologisch und konventionell angebauten Kartoffeln (*Solanum tuberosum, var. Desirée*) unter Verwendung eines Post-Bisulfit-Adapter-Tagging Whole-Genome Bisulfit-Sequencing (PBAT-WGBS) Protokolls wurde im Rahmen dieser Dissertation untersucht. Dieser neue Ansatz nutzt die umweltabhängige Anlagerung von Methylgruppen an Cytosine innerhalb der DNA, um nach Unterschieden zwischen Proben zu suchen, die unter verschiedenen Anbausystemen gewachsen sind. Diese Methode wurde an Proben aus dem DOK-Feldversuch (Agroscope und Forschungsanstalt für biologischen Landbau, Frick, Schweiz) getestet. Die Proben wurden unter Feldbedingungen in drei Feldwiederholungen und in zwei Jahren mit der selben Sorte und den selben Saatknollenchargen für die konventionellen und bio-organischen Parzellen angebaut. Dabei wurden die bio-organischen Parzellen Kupferapplikationen zur Pilzbekämpfung und organischer Düngung unterzogen, während die konventionellen Parzellen mit synthetischen Fungiziden behandelt und mit mineralischen sowie organischen Düngemitteln gedüngt wurden.

Die vorliegende Arbeit ist in drei Teile unterteilt, wobei der erste und zweite Teil bereits in internationalen, peer-reviewed Journalen publiziert sind. Für den dritten Teil ist eine Publikation der Ergebnisse geplant.

Im Rahmen eines peer-reviewed publizierten Review-Artikels wurden die derzeit im Bereich der DNA-Methylierungsanalyse angewandten Schritte, Methoden und Datenanalysen beschrieben. Für unseren Versuch wurde PBAT-WGBS aufgrund der Zuverlässigkeit, der Auflösung bis auf Einzelbasenebene und der Fähigkeit, geringe Mengen an DNA als Input zu verwenden, ausgewählt.

Um die entsprechenden Datensätze der PBAT-WGBS-Analysen auszuwerten, wurde eine zuverlässige Pipeline zur bioinformatischen Auswertung der Methylierungsmuster benötigt. Diese wurde in snakemake, einem Pipeline-Analyse-Tool, implementiert. Das Alignment wurde, nach einem intensiven Qualitätsbenchmarking von Mapping-Tools, mit Bismarkbwt2 durchgeführt. Das Ergebnis dieser Benchmark-Studie mit 8 Sequencing-Read-Mapping-Tools und verschiedenen Parametersatz-Kombinationen wurde ebenfalls in einem peer-reviewed Artikel veröffentlicht.

Schließlich wurden differentiell methylierte Cytosine (DMC) und differentiell methylierte Regionen (DMR) in beiden untersuchten Jahren gefunden, die mit Stoffwechselwegen, Transportmechanismen und dem Pflanzenhormonsystem in Verbindung stehen. Eine DMR in einem einzelnen Jahr war mit ATOX1 assoziiert. Dies deutet auf eine Entgiftung von Kupfer in Bio-Kartoffeln hin, muss aber in weiteren Studien bestätigt werden.

Die Methode wurde exemplarisch in Kartoffeln aus Bioanbau sowie aus konventionellem Anbau angewandt. Neben dieser Anwendung bietet sie die Möglichkeit Lagerungs-, Transport-, Verarbeitungs- sowie weitere Anbaubedingungen in einer Vielzahl an Produkten zu analysieren. Daher wird sie in der zukünftigen Produktauthentifizierung eventuell von Bedeutung sein.

Zusammenfassend konnte die Zuverlässigkeit der angewandten Methode nachgewiesen werden.

1. Introduction

Distinguishing analytically between organic and traditionally/conventionally grown plants is of high importance to assure consumers´ confidence in agricultural production systems as well as labelling of food and other plant-based products. Furthermore, there is need for such a method for peer-to-peer trading within the value chain for verification of the product quality and for food fraud detection in case of concerns.

The term "organic agriculture" describes highly different agricultural land use practices. We will focus on the EU standard. This includes mainly the absence of synthetic plant protection and fertilization and a more diverse crop rotation.

In the following, I briefly summarize the methods that have been tried to distinguish between organic and conventional food samples and their respective results. Further, a new single-technique-approach will be proposed, which has the potential to enrich the scientific discussion about this topic.

Coffee (De Nadai Fernandes, Elisabete A. et al. 2002), peas, onions (Gundersen et al. 2000), winter wheat, barley, faba beans and potatoes (Laursen et al. 2011) have been tried to discriminate using element analysis with mostly good success, whereas the involvement of non-essential elements seems to enhance the right classification. Because of the absence of synthetic fungicides and the use of organic matter instead of mineral fertilizers, the arbuscular mycorrhizal fungi (AMF) spore abundance is significantly higher in organic fields compared to conventionally managed ones (Oehl et al. 2004). This enables organically grown plants to obtain via symbiosis slightly more calcium, copper, zinc and rubidium (Kelly and Bateman 2010) and leads, combined with a lower growth rate and dry matter content (Alaru et al. 2014) to higher concentrations of these elements than in conventional soils (Gosling et al. 2006). Also decreasing of trace elements such as manganese has been reported for example in tomatoes (Kelly and Bateman 2010) and maize (Kothari et al. 1991).

The bottleneck of this method is that the variability of concentration of major and trace elements depends mainly on the variety of species and, therefore, could not be generally used as a reliable marker for organic food (Ordóñez-Santos et al. 2011). Also the growing site with different weather conditions and soil types, the crop cycle and the physiological age of the harvested product as well as the highly varying nutrient availability in "organic" soils influences the chemical composition of elements in plant products. The trade market for most trading goods does not differ between species and the mentioned growing conditions. The number of possible markers is restricted because of the number of elements, which seems to be too low to see clearly separable patterns if different species and years are included in experiment design. The farming practice is just one trait that could change the amount of major trace elements and could be overlaid by other traits (Magkos et al. 2003; Rosen and Allan 2007).

Stable isotope analysis is another promising analytical method to distinguish conventional from organic food products. The δ ^{15}N value of atmospheric nitrogen is defined as zero ($^{15}N/^{14}N$ ratio = 0.00368). Also the value of mineral nitrogen fertilizers, which is used in conventional agriculture, derived by Haber–Bosch process is at about this ratio. A combination out of preferential volatilisation of ^{14}N through bacteria activity and higher ^{15}N content of manure (Bateman and Kelly 2007), both because of enzymatic discrimination of the heavier ^{15}N, leads to higher $^{15}N/^{14}N$ ratio in organically fertilized soils. The ratio within the soil is known to represent the range of the $^{15}N/^{14}N$ ratio into the grown plants (Choi et al. 2003). Therefore, it has been concluded that the $^{15}N/^{14}N$ ratio in conventionally fertilized plants should be lower than in plants grown with compost and manure. Many studies have explored the differences of stable isotope ratio, especially of nitrogen with whole tissue mass spectrometry, in products obtained from organic or conventional agricultural systems. This has been done for various crops and vegetables like grain crops (Kohl et al. 1973), endive, rocket, leek, potato, chicory, sweet pepper, garlic, onion (Sturm and Lojen 2011), carrot, tomato (Bateman et al. 2005), lettuce, cabbage, Chinese cabbage (Schmidt et al. 2005) and citrus fruits (Rapisarda et al. 2005).

In conclusion, type of fertilization could be monitored in case that exclusively organic fertilization has been chosen. Because of many limitations, a combination out of several other methods has been proposed after detailed literature review (Laursen et al. 2014). Using different fertilization is not the only feature defining organic agriculture. Differences in plant protection cannot be inspected. Further it is also permitted to use manure and compost in conventional systems, which complicates the identification process. Even if the authentication process is improved via using multi-isotope analysis (Laursen et al. 2013) or compound-specific analysis instead of whole tissue analysis of stable isotope ratios, as it has been done for amino acids in wheat and durum (Paolini et al. 2015), some challenges remain. Also within the *Authenticfood* project, a part of Core Organic 2 – a large EU-funded project on food safety, the authors finally recommended to focus on compound-specific stable isotope analysis (Husted 2015).

Legumes and slower growing plants such as potatoes could not be addressed by this method (Rogers 2008). Other authors concluded a threshold value of 4.3‰ δ ^{15}N for potatoes using a controlled field trial, whereupon still 15% of conventional samples have been misclassified (Camin et al. 2007). The failure of this method could be explained since the timing of fertilization as well as the age and the type of the collected plant organ also varies the ^{15}N/^{14}N ratio (del Amor et al. 2008). Another point that causes trouble is the biologically fixed atmospheric nitrogen by legumes and other cover crops in organic agriculture, which is lowering the δ ^{15}N value of the soil similar to mineral fertilizers do and, therefore, leads to confusing results (Laursen et al. 2013).

X-ray spectroscopy and infrared spectra analysis have been tried to answer different questions of food quality observation (Putzig et al. 1994). The publications on this topic showed correct classification rates between 74% and 93% for infrared spectra analysis combined with multivariate statistics of organic red and white wines (Cozzolino et al. 2009), rates up to 91% for asparagus (Sánchez et al. 2013) and a shift for potassium peak in the x-ray spectra of tomato and coffee samples (Bortoleto et al. 2008). This stands in contrast to the observed results of element analysis where no significant differences in potassium content could be spotted. At this moment, the results are not accurate enough to clearly identify organic products while the basis of literature on spectroscopy-based methods as well is too low.

The basis of published literature on copper chloride crystallization application in terms of analytical identification of organic products is also not well established at this moment. For a long period of time, copper chloride crystallization with plant extract solution on glass dishes was not accepted due to difficult validation and reproducibility of the obtained crystallization pictures because the analysis of pattern had to run visually. Busscher et al. (2010) described a self-learning, computerized analysis tool to detect pattern in the scanned crystallization images that encode for organic agriculture. Szulc et al. (2010) applied this tool to winter wheat samples and got classification rates ranging from 69 to 96% for single years. But if two different years where combined for training of the tool, the model was not suitable because of too high variability between the years. The classification rates fluctuated between 56 and 59%. A study with visual and computerized picture analysis as well got highly varying results and only for 2003, a rate of 100% correct classification for the organic samples could be achieved (Kahl et al. 2008).

An alternative approach to differ between organically and conventionally produced plants could be the different application of pesticides. A meta-analysis about pesticide residues in food samples showed a significant increase of some agents in conventional vegetables and potatoes (Hoefkens et al. 2009). But for successful separation of harvested products, this method is not suitable because it relies on just one agronomical variation - the prohibition of chemical pesticides. Via persistent soil contamination, spray drift by wind and cross-contamination during storage, processing and packaging, pesticide residues could also be found in organic products (Guignard et al. 2015). Furthermore, it does not guarantee the compliance of other ecological farming practices if there could not be found any pesticide residues.

Other approaches looked at ingredients such as phenolic or volatile organic compounds, proteins and metabolites as possible fingerprints with different success. Compared to conventionally managed vegetables, for organically produced tomatoes, a higher amount of flavonoids has been reported (Chassy et al. 2006; Mitchell et al. 2007) as well as for qing-gen-cai (chinese cabbage), spinach, welsh onion and green bell pepper (Ren et al. 2001). There is also a shift in the content of anthocyanin in organic blueberries (Wang et al. 2008). Divergences in total phenolic concentration have been published for organic eggplants (Raigon et al. 2010), beets (Rossetto et al. 2009), peaches, pears (Carbonaro et al. 2002) and marionberries (Asami et al. 2003). Processed food, such as organic ketchup (Vallverdu-Queralt et al. 2011) and wines (Vrček et al. 2011), seems to show higher phenolic content if it is derived from organic agriculture.

Studies which report no significant differences in concentrations of phenolic compounds between organically and non-organically produced foods are also available for example for strawberries (Häkkinen and Törrönen 2000), eggplants (Luthria et al. 2010), beetroots (Kazimierczak et al. 2014), grapes and wines (Mulero et al. 2009, 2010). Because of these reports, it is questionable to use the analysis of phenolic compounds for practical application. The variation of phenolic compounds with maturity stage and pathogen attack makes it further challenging because organic crops generally have a longer growing period than conventional crops and, therefore, they have a longer period to build up their phenolic pool (Jeffery et al. 2003). Also the chemical composition of vegetables and fruits is mainly influenced by choice of cultivar (Chassy et al. 2006).

One study examined the variability of volatile organic compounds (VOC) in combination with sensory testing for organic and conventional tomatoes (Muilwijk et al. 2015). In this case, the VOC fingerprint was mainly affected by location and only in a second step by agricultural system whereas the cultivar had a minor effect.

As expounded above, several methods have been tried and non could sufficiently identify organic plant products. Therefore, there is still need for a reliable approach that can enrich the scientific discussion about this topic and approach the given aim.

Epigenetics analyses and especially DNA methylation analyses is capable to fulfill this task as, driven by environmental conditions, patterns on top and beside the DNA are adapted to react on different environments.

To minimize redundancy, for more information about epigenetics and DNA methylation please find detailed information in chapter 3.2, 4.2 and 5.2.

2. Research Question

The following work addresses the the question of analytical identification of organically grown plant samples. The overall aim has been to test whether it is possible to differentiate between conventional and organic plant samples based on DNA methylation analysis. Thereby, using Bisulfite-Sequencing, unmethylated Cytosines are converted by Bisulfite to Uracil and subsequently to Thymines and can be identified after mapping to a reference genome. Methylated Cytosines are "protected" from conversion by Bisulfite and remain as Cytosines.

For the purpose of differentiation, patterns as "epigenetic fingerprint" were used, which might be taken as reliable markers to unambiguously identify organic food.

The following hypotheses (H) were tested in detail:

> H1 The agricultural practice under which a plant has been grown could be monitored using analysis of DNA methylation status in harvested products like potato tubers.
>
> H2 Based on transcriptome and metabolome studies, methylation differences were expected due to plant defence, hormone and metabolic pathway related genome regions as well as in transposal regions.
>
> H3 The epigenetic variation of some DNA methylation islands between the years is lower than between the agricultural systems. Therefore, DNA methylation fingerprinting could be used as biomarker to confirm the food production system.

It is necessary to show every difference in terms of methylation status during the first run, even in untranslated regions, because these parts of the genome could affect transcriptional gene activity due to methylation-derived differences in chromosome structure, too. However, using the technique of Whole-genome Bisulfite Sequencing (WGBS), it could be benefited from the large amount of bases. Therefore, a much wider pool of potential biomarkers is available than for example in transcriptome or metabolic studies because every Cytosine could be differently methylated. The following down-scaling process will decrease the number of differently methylated sites every time the next factor of variation is included. The variability between different years will be used as a major variability factor to look for reliable DNA methylation differences attributed to organic agriculture.

The topic will be addressed in three independent chapters including the planning process of Whole-genome-bisulfite-sequencing experiments (chapter 3, Grehl et al. 2018), the benchmarking of WGBS alignment software (chapter 4, Grehl et al. 2020) and the exemplary analysis of DNA methylation patterns in organic and conventionally grown potato samples (chapter 5, submitted for publication). Parts of this dissertation are already published (Grehl et al. 2018; Grehl et al. 2020). For the third part, the submission as manuscript for publication is planned.

3. How to Design a Whole-Genome Bisulfite Sequencing Experiment (Review)

Published in: Epigenomes, 2 (4), p. 21, 2018; https://doi.org/10.3390/epigenomes2040021

Special Issue: Plant Epigenetics

Claudius Grehl [1,2*], Markus Kuhlmann [3], Claude Becker [4], Bruno Glaser [2] and Ivo Grosse [1,5]

[1] Institute of Computer Science, Martin Luther University Halle-Wittenberg, Von Seckendorff-Platz 1, 06120 Halle (Saale), Germany; ivo.grosse@informatik.uni-halle.de

[2] Institute of Agricultural and Nutritional Sciences, Martin Luther University Halle-Wittenberg, Soil Biogeochemistry, von-Seckendorff-Platz 3, 06120 Halle (Saale), Germany; bruno.glaser@landw.uni-halle.de

[3] Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Corrensstraße 3, 06466 Gatersleben, Germany; kuhlmann@ipk-gatersleben.de

[4] Gregor Mendel Institute of Molecular Plant Biology, Austrian Academy of Sciences, Vienna Biocenter (VBC), Dr. Bohr-Gasse 3, 1030 Vienna, Austria; claude.becker@gmi.oeaw.ac.at

[5] German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Deutscher Platz 5e, 04103 Leipzig, Germany

* Correspondence: claudius.grehl@informatik.uni-halle.de

3.1. Abstract

Aside from post-translational histone modifications and small RNA populations, the epigenome of an organism is defined by the level and spectrum of DNA methylation. Methyl groups can be covalently bound to the carbon-5 of cytosines or the carbon-6 of adenine bases. DNA methylation can be found in both prokaryotes and eukaryotes. In the latter, dynamic variation is shown across species, along development, and by cell type. DNA methylation usually leads to a lower binding affinity of DNA-interacting proteins and often results in a lower expression rate of the subsequent genome region, a process also referred to as transcriptional gene silencing. We give an overview of the current state of research facilitating the planning and implementation of whole-genome bisulfite-sequencing (WGBS) experiments. We refrain from discussing alternative methods for DNA methylation analysis, such as reduced representation bisulfite sequencing (rrBS) and methylated DNA immunoprecipitation sequencing (MeDIPSeq), which have value in specific experimental contexts but are generally disadvantageous compared to WGBS.

Keywords: WGBS; coverage; library; DNA methylation; 5mC; epigenome; epigenetics

3.2. Introduction

Aside from post-translational histone modifications and small RNA populations, the epigenome of an organism is defined by the level and spectrum of DNA methylation (Law and Jacobsen 2010). Cytosine methylation can occur as 5-methylcytosine (5mC) or 5-hydroxymethylcytosine (5hmC) (Wyatt and Cohen 1953). Its key roles are in embryonic development regulation, genome imprinting, silencing of transposons, cell differentiation, X-chromosome inactivation, and general transcriptional gene regulation (Hackett and Surani 2013; Zhang et al. 2018). The addition of methyl groups onto DNA bases generally represses gene expression and therefore acts as one of several transcription control mechanisms.

Whole genome bisulfite-sequencing has become the gold standard to detect DNA methylation patterns because of its single-base resolution and the possibility to cover entire genomes. Other approaches use either pre-selection and single-base resolution (Sun et al. 2018) or region-based approaches in whole genomes (Bock et al. 2010; Aberg et al. 2017) and, therefore, do not deliver the full spectrum and detail of DNA methylation patterns.

Since the development of bisulfite-treated DNA sequencing in 1992 by Frommer et al., the application of high-throughput sequencing has been especially useful in facilitating the reliable detection and analysis of methylation patterns in several organisms, tissues, and cell types. Huge steps in the field of DNA methylation analysis were the first WGBS of *Arabidopsis thaliana* (Cokus et al. 2008; Lister et al. 2008), and the first human methylome sequencing in 2009 by Lister et al.. During the bisulfite reaction of DNA, unmethylated cytosine is converted to uracil (Figure 1). This occurs via cytosinsulphonate, a further hydrolytic deamination step to uracilsulphonate, and, finally, a desulphonation step to uracil (Figure 1). However, 5mC and 5hmC are inert to this chemical conversion.

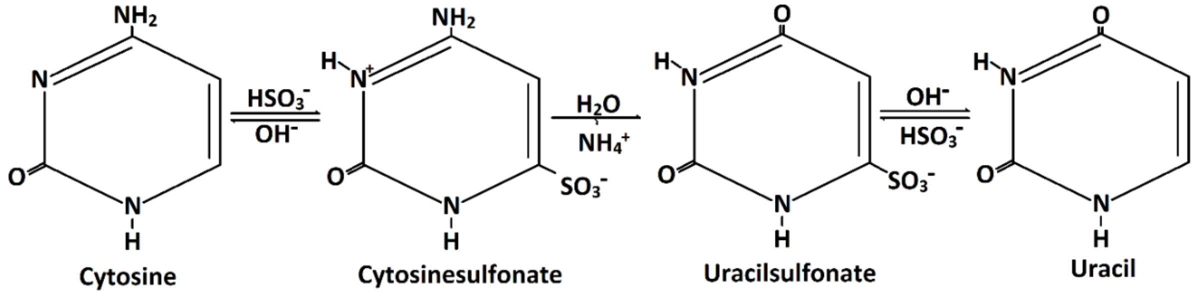


Figure 1: Principle of the bisulfite-mediated conversion of cytosine to uracil.

The subsequent polymerase chain reaction (PCR) translates uracil into thymine. During sequencing, this base pair shift causes cytosine/thymine-polymorphism, which can be quantified and interpreted as proportion of the original methylation at a specific site through the comparison of reads with the original strand or a reference genome (Figure 2).

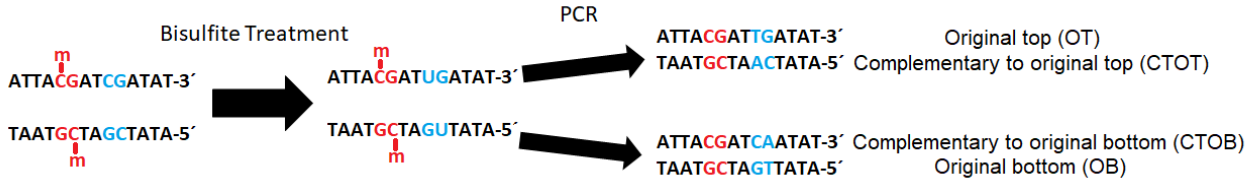


Figure 2: Exemplary DNA double strand with methylated (red) and unmethylated (blue) CpG-site (cytosine-phosphate-guanine-dinucleotide) before and after bisulfite application and polymerase chain reaction (PCR). Methylated cytosine is not affected by bisulfite, whereas unmethylated cytosine is converted to uracil and further on to thymine during PCR in the original top strand, and to adenine in the complementary top strand.

A disadvantage of WGBS is that bisulfite sequencing cannot distinguish between 5hmC and 5mC. The 5hmC is described as the dynamic DNA modification associated with the DNA demethylation process (Shi et al. 2017) in animals. This has to be taken into account for organisms/cell types with substantial amounts of 5hmC, such as the cerebellum cells of Parkinson disease patients (Stöger et al. 2017). In plants, 5hmC is of minor relevance (Erdmann et al. 2014), as it occurs only in very low amounts and may merely serve as an intermediate product during demethylation (Wang et al. 2015).

A second disadvantage arises from the C–T base switch, particularly when analyzing non-reference strains or mutagenesis populations, as it can become impossible to distinguish between bisulfite-induced deamination of unmethylated cytosine on the one hand and true C-to-T single nucleotide polymorphism (SNP) on the other.

In plants and some animals, large parts of the genome are methylated depending on the genome size, genome duplications, and the control of most of these duplicated parts. In most animals, cytosine carries methylation predominantly in the CG context accumulated in CpG islands for gene regulation, often near transcription starting sites (Deaton and Bird 2011). In contrast to that, plant genomes have additional CHG and CHH methylation (H = adenine, thymine or cytosine). These last two correlate with the suppression of transposon activity (Zakrzewski et al. 2017), mobile genetic elements that can change their position within the genome. While CG and CHG methylation are symmetric, i.e., the palindromic cytosine on the complementary strand is usually also methylated, CHH methylation is asymmetrical (Selker and Stevens 1985). The proportion and context of cytosine methylation ranges from <1 % total cytosine methylation during the early development of *Drosophila melanogaster* (Lyko et al. 2000) to 74.7 % CG, 69.1 % CHG, and 1.5 % CHH methylation in *Picea abies* (Ausin et al. 2016).

DNA methylation patterns are often tissue-specific (Zhou et al. 2017), environment-altered (Flores et al. 2013; Kumar et al. 2017), and can be stably inherited to subsequent generations (Finnegan et al. 1996; Niederhuth and Schmitz 2014).

Because of the importance of DNA methylation as transcriptional regulator, its analysis plays an increasing role in integrative -omics studies, ecological examinations, and medical research. The aim of this publication is to facilitate the launch of WGBS experiments. Therefore, we will give a brief overview of WGBS experiment planning by discussing different types of WGBS library preparation protocols, options for sequencing technology, and bioinformatics data analysis.

3.3. Whole Genome Bisulfite-Sequencing Preparation

The steps, prior to sequencing, needed for WGBS library preparation are: DNA extraction, DNA fragmentation, DNA repair, adapter ligation, bisulfite treatment, PCR amplification (Figure 3) including size selection, and quality control at different points in time. These can be arranged in two general ways. The pathway shown on the left side of Figure 3 is a pre-bisulfite protocol. Because the bisulfite treatment takes place after the adapter ligation, these protocols require methylated adapters that will be inert to the treatment. In post-bisulfite protocols, like a post-bisulfite adapter tagging protocol, DNA fragmentation is facilitated through the bisulfite treatment itself and is followed by DNA end repair and adapter ligation or random priming including streptavidin-coated magnetic beads (Miura et al. 2012). The advantage of the latter is that less DNA is needed as less adapter-ligated DNA fragments are destroyed by the harsh conditions of bisulfite treatment (Peat and Smallwood 2018). A disadvantage of this kind of library preparation is the slightly higher computational power needed to bioinformatically process these non-directional reads. Usually, over 90 % of the adapter-ligated, intact fragments are destroyed in pre-bisulfite protocols, even under ideal conditions (Kint et al. 2018). In the end, the bisulfite conversion step leads to the conversion of cytosine to uracil, which is much more susceptible to degradation (Tanaka and Okamoto 2007). The decline in the amount of DNA could be compensated by PCR amplification, but should be done with as few cycles as possible to yield the lowest possible PCR bias. Minimum two PCR cycles are needed because the former unmethylated cytosines have to be rewritten from uracil to thymine. The used polymerase must have the ability to read uracil. Therefore, most polymerases with proofreading and repair functions, except of Pfu Turbo Cx, cannot be used. Another common polymerase for this purpose is KAPA HiFi Uracil+. Furthermore, highly PCR-amplified bisulfite-converted libraries may be unbalanced, as highly methylated, and fragments with high C content post-conversion are easier to amplify and, therefore, will be overrepresented in the final library (Ji et al. 2014).

The fragment length depends on the sequencing technology. The desired read length could be adjusted after DNA shearing by gel selection, bead selection, or BluePippin (Sage Science), an automated DNA and RNA size selection system. Depending on the final sequencing platform, fragments should have a length of 200–400 base pairs (without adapters; see Sequencing technology). Classic gel selection has to be made, for example, with SYBR Gold stain instead of ethidium bromide to avoid DNA degradation and contamination. The desired fragment length could be easily cut out of a 1 % agarose gel after DNA fragmentation if a DNA standard has also been run on the same gel. This process was automated in BluePippin by the application of a gel cassette. However, most library preparation kits today use magnetic beads for size selection.

To calculate the bisulfite conversion rate, unmethylated reference DNA has to be added to the library prior to the bisulfite treatment, at a ratio of 0.1–0.5 % (w/w) of the total DNA. Lambda phage DNA is most commonly used for this purpose. It is completely unmethylated and is, therefore, expected to be converted at every cytosine position. In plants with assembled chloroplast genomes, the unmethylated chloroplast DNA, which is contained to some degree in all genomic DNA extracts, can be used instead of or on top of lambda. In non-plant species it is also possible to use a non-CG methylation context to control the conversion efficiency (Warnecke et al. 2002).

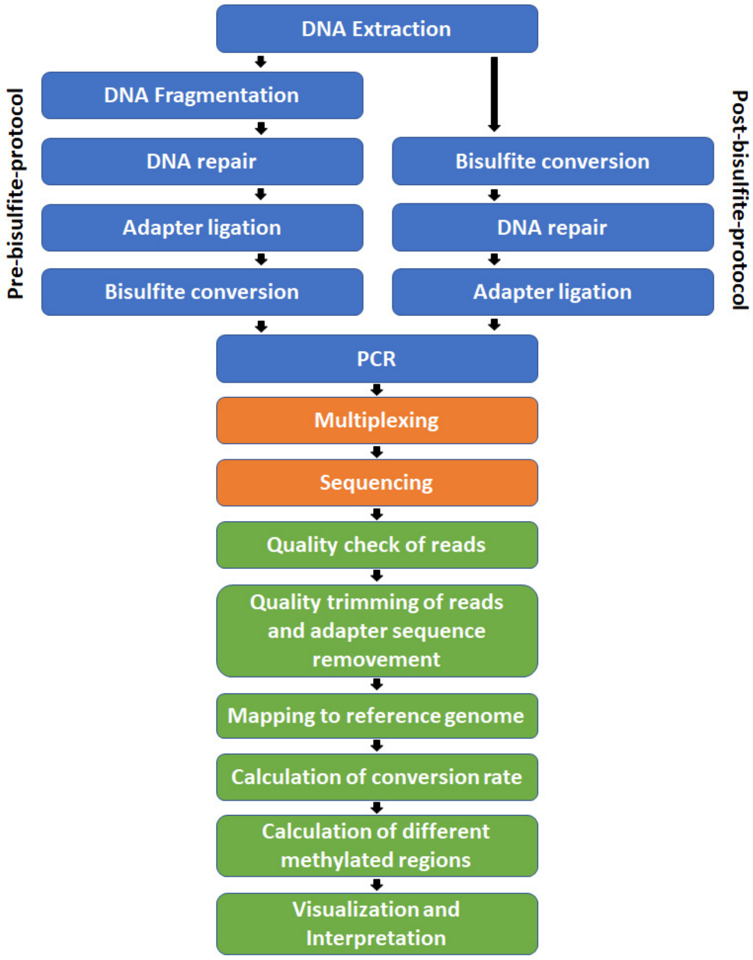


Figure 3: Flowchart of the core steps in WGBS experiments; blue = library preparation split into an exemplary pre-bisulfite-protocol and a post-bisulfite-protocol, orange = sequencing, green = bioinformatic analysis.

Several protocols for WGBS library preparation have been proposed using kits from different companies. The decision depends heavily on the experience of the individual laboratory and, therefore, is highly subjective. In general, a fast and routine working process is of importance to avoid batch effects and the introduction of a bias due to amplification. This should be avoided by working with amplification-free libraries (McInroy et al. 2016; Olova et al. 2018).

General examples for library preparation kits are EpiTect Plus (Qiagen, Venlo, Netherlands), EZ DNA Methylation-Gold Kit (Zymo Research, Irvine, United States of America), or Imprint® DNA Modification Kit from Sigma-Aldrich/Merck (St. Louis, United States of America,) for bisulfite conversion and Accel-NGS® Methyl-Seq DNA Library Kit (Swift Bioscience, Ann Arbor, United States of America), TrueSeq Nano (Illumina, San Diego, United States of America), or SureSelectXT Methyl-Seq Target Enrichment System (Agilent, Santa Clara, United States of America) for adapter ligation.

The cost of a WGBS library is currently in the range of 50–240 €/sample (2018), depending on the library preparation method, supplier, and conversion kits.

3.4. Sequencing Technology

Aside from Roche, Thermo Fisher Scientific Inc., Beckman Coulter, and Pacific Biosciences, the market leader with more than 75 % of the market share in terms of next generation sequencing technology is Illumina (Global Next Generation Sequencing Market Assessment & Forecast), offering several platforms for different applications. For detailed information see the Illumina platform website (https://emea.illumina.com/systems/sequencing-platforms.html).

Due to DNA degradation during the bisulfite treatment, the production of long fragments for sequencing is limited. The longer the fragment, the more likely is a fracture of the strand due to bisulfite treatment. Therefore, the advantages of long fragment generating systems, such as more reliable mapping of reads, currently do not apply. Benchtop solutions are partly not suitable because of the much higher price per giga base (Gb) of produced sequence. A comparison of the most promising systems currently used, such as HiSeq, HiSeqXTen, NovaSeq, and PacBio, is discussed in the following section.

For high-throughput sequencing, several systems exist, such as HiSeq2500, HiSeq3000 and HiSeq4000. They differ in terms of run time, maximum read length, and read output, but also in their acceptance of certain types of libraries, their potential to deal with low complexity libraries, and their quality score output. Compared to HiSeq3000 and HiSeq4000, the HiSeq2500 system can deal with low sequence diversity libraries like amplicon or WGBS libraries because of an optimization of the cluster calling algorithm. Nevertheless, the other systems could also be used if the libraries had been correctly multiplexed with other libraries with balanced base proportions at each read position or minimum 20 % phi X DNA (see Multiplexing).

HiSeq sequencing costs are around 60–90 €/Gb (2018) depending on the read length and the option of paired- or single-end sequencing (see Figure 4). Theoretically, paired-end sequencing offers the potential to map deeper into repeat-regions because two reads are generated, one from either side of the fragment. This has, so far, not been investigated systematically but seems to depend on the mapper used (Tran et al. 2014). Single-end sequencing is lower in price but faces lower unique mapping rates in repeat-rich regions. For paired-end sequencing, it has to be ensured that the fragment length is more than two times as large as one single read. Otherwise, information is generated twice in the same fragment and the information/price ratio worsens. It has been shown that the error rate could be reduced and the sensitivity enhanced by usage of paired-end bisulfite sequencing (Tsuji and Weng 2016). Therefore, we recommend using paired-end sequencing.

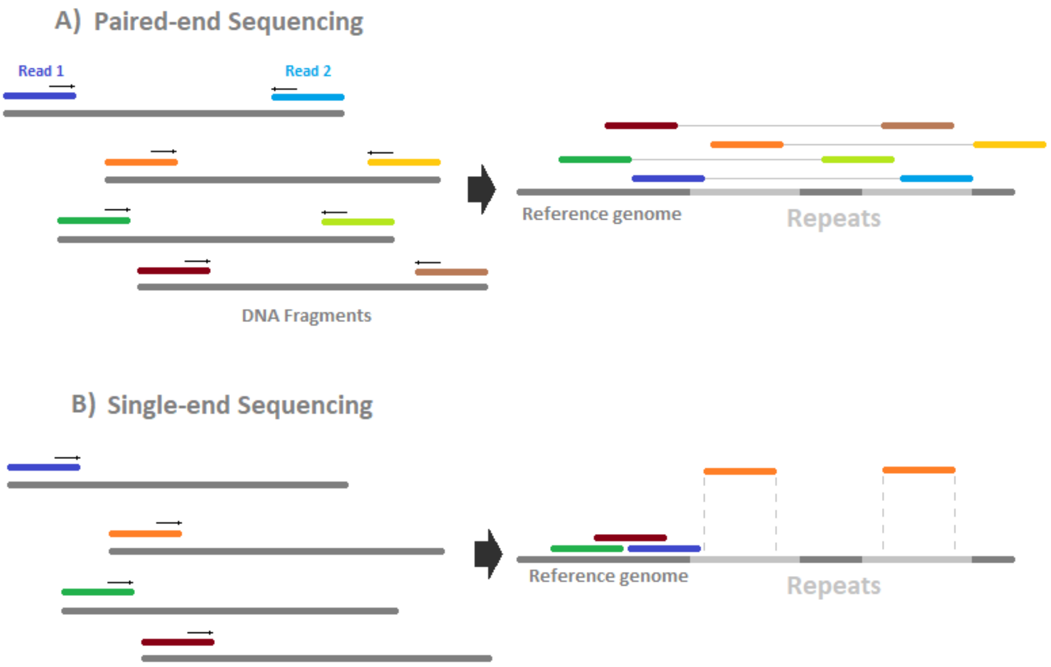


Figure 4: Principle of paired-end and single-end sequencing and mapping on a reference genome with repeats. (A) Paired-end sequencing: Single-stranded DNA fragments are sequenced by synthesis from both sides, which generates two reads per fragment. During mapping, within a defined distance of a uniquely mapped read, the corresponding second read is searched and mapped. This allows a reliable mapping of more reads in repeat-rich genomes for short repeated parts of the DNA. (B) Single-end sequencing generates reads only from one side of the fragment. As the chemistry behind single-end sequencing is cheaper, the same base pair output is generated in a more cost-effective way. Short DNA fragments are covered with a greater efficiency compared to paired-end reads. Repeat-rich genomes face the challenge of more multiple mapped reads, as shown for the orange read, compared to paired-end sequencing.

A less cost-intensive approach was set up in the HiSeqXTen (Nair et al. 2018), by applying a combination of ten HiSeq systems. Formerly released only for human samples, the system has been opened to other large-scale sequencing experiments of non-human samples (Illumina press release, October 6, 2015). Dependent on the sequencing facility, prices as low as circa 14 €/Gb (2018) were offered.

NovaSeq 6000 is the most recent production scale sequencer and uses the two-color chemistry already shown in the NextSeq system. By combining two nucleotide-binding fluorescence dyes, a four-letter code can be accomplished. A disadvantage of this method is that no signal (zero) codes for one of the four bases within the sequencing process. This lowers the accuracy of the data but offers a faster sequencing process to a much lower price, in the range of HiSeqXTen systems (Raine et al. 2018). Only two wavelength-filtered images of the flow cell need to be computed compared to four images in a four-color chemistry system such as in HiSeq. For WGBS libraries, this technique is not recommended, as a higher GC bias is introduced by bisulfite conversion that could not be tolerated by two-color-chemistry.

Single molecule, real-time sequencing (SMRT-S), such as PacBio and NanoPore sequencing, has shown reliable results for the direct detection of adenine methylation in bacteria based on signal delay of fluorescence-labelled nucleotides (Flusberg et al. 2010) and for 5mC analysis of human DNA (Simpson et al. 2017). However, compared to HiSeq or NovaSeq, costs are high (>150 €/Gb). Furthermore, the genome read coverage of such experiments has to be much higher to compensate for the base call error rate of 5–10 % while Illumina HiSeq sequencing yields circa 0.0034–1 % false base calls (Escalona et al. 2016). If these bottlenecks could be solved, the possibility of long reads and, therefore, the facilitated mapping of repeat-rich regions, and the removal of bisulfite treatment from the protocol would be highly advantageous.

3.5. Multiplexing

To reduce the risk of losing data due to experimental flaws during sequencing, to reach high coverages and read numbers, several samples and types of libraries should be sequenced in combination to yield a high sequencing depth on multiple lanes. A global, equimolar library should be produced based on the quality and quantity (Qubit and/or qPCR measurement) of the DNA sample libraries. The intended sequencing depth or genome coverage is one of the most significant cost factors for WGBS experiments, aside from the number of replicates and the genome size. A sufficient number of reads per covered genome region defines the quality and the statistical power of the downstream analyses. False-positive base calling could be detected more reliably and methylation levels could be determined more accurately at sufficient sequencing depth.

Balancing the coverage and the number of replicates is one of the most important steps within the planning process of WGBS. Ziller et al. (2015) recommended a coverage of 15 fold and three replicates. Beyond that, money would be better spent strengthening the statistical power by increasing the number of statistically independent biological replicates. The theoretical coverage has to be estimated based on the number of homologous chromosomes of the species, the amount of repeats, and the expected degree of heterozygosity, as the level of methylation could differ between the paternal and maternal alleles. For repeat-rich genomes, a coverage of at least 20 fold should be taken into account.

The estimated genome size—not the size of the actual reference assembly—has to be considered for the calculation of total necessary data output as most of the more complex reference genomes assemblies are incomplete, including many scaffolded parts.

Particularly for bisulfite-treated samples, it is important to multiplex with non-bisulfite-treated libraries, as sequencing technology currently tends to perform worse with extremely unevenly distributed base proportions (e.g. GC-rich or AT-rich libraries). Depending on the rate of methylation, due to bisulfite treatment and the subsequent PCR, generated bisulfite libraries contain a shift in base proportions. The proportion of adenine/thymine is enlarged while the guanine/cytosine proportion is reduced because unmethylated cytosines are converted to thymines via the uracil-intermediates (Figure 2). For the second paired-end read the proportion of complementary bases shifted because of the bridge amplification during sequencing (Figure 5). So far, no batch effect between lanes has been reported in literature.

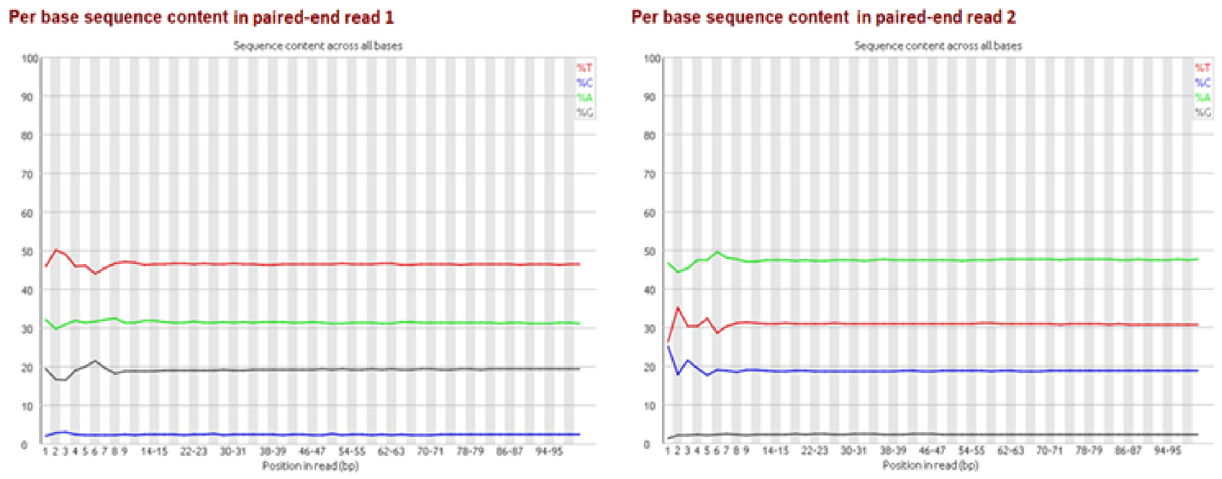


Figure 5: Base proportion per position over all reads of a whole-genome bisulfite library of soybean (Glycine max) seed coat with a cytosine methylation proportion of circa 11 %, paired-end sequencing 2 × 100 bp, red line: thymine proportion, green line: adenine proportion, black line: guanine proportion, blue line: cytosine proportion; left: first paired-end read, right: second paired-end read after bridge amplification. The base proportion shift could be explained with the bisulfite conversion and the subsequent PCR, namely the enrichment of thymines and the reduction of cytosines. Graphic by FastQC, a tool for quality measurement of next-generation sequencing data.

3.6. Bioinformatics Tools and Benchmarking

An informative overview of the required bioinformatics steps for WGBS after sequencing has been published by Shafi et al. (2017). Quality check, adapter removal, alignment, methylation calculation, and calculation of differentially methylated regions (DMRs) are the core steps within the bioinformatics pipeline of WGBS experiments (Figure 3) and could be done by several open-source tools.

As every organism differs in reference genome assembly quality, amount of repeated regions, and homogeneity of base proportion, the applied tools have to be evaluated, ideally by benchmarking them on simulated datasets based on the separate reference genome for every species. The mapper for bisulfite-reads and the DMR caller should be included in such a comprehensive survey.

The alignment tools themselves differ in terms of the algorithms and the strategy used, e.g., three-letter or wild-card alignment, as well as the output file format and content. Some tools report only uniquely mapped reads, whereas others also include multiple-mapped or discarded reads in their output. The alignment tools have to cope with the four different, uncomplimentary strands as a result of the combination of bisulfite treatment and PCR, as shown in Figure 2. The performance of alignment tools differs, depending on the size and amount of repeats in the genome, the coverage, and the sequencing technology used. Hence, for every project the best tools have to be identified and combined to form a data analysis pipeline.

The main difficulty in benchmarking studies is that a known truth has to be generated for multiple class hypothesis testing by the application of simulated datasets. A comprehensive study on simulated and real human lung tumor tissue rrBS data was performed by Sun et al. (2015) using the simulation tool RRBSsim. Here, the tools bwa-meth and BS-Seeker2 showed reliable results for sensitivity, precision, and speed in the mapping of simulated data. For WGBS, no such study was available but, for example, the tool "Sherman— bisulfite-treated Read FastQ Simulator"could be used for the simulation of WGBS datasets based on a given reference genome with parameters like number and size of reads, paired-end or single-end sequencing, conversion rate, number of SNPs, and error rate. Other often used mapping tools are Bismark (Krueger and Andrews 2011), BSmap (Xi and Li 2009), and Segemehl (Otto et al. 2012).

For DMR calling, the available approaches differ in terms of their capability to consider replicates and coverage dependencies, the type of boundary estimation and the statistical tests, such as logistic regression, binary segmentation, and beta-binomial-based approaches (Escalona et al. 2016).

The definition of a known truth and the definition of DMRs (e.g., size, coverage, and grade of methylation difference) are also of importance. A small subset of differentially methylated cytosine (DMC) detection tools have been included into a benchmarking analysis by Wreczycka et al. (2017) showing moderate results for methylKit. Better results have been published for metilene (Jühling et al. 2016) or Defiant (Condon et al. 2018). In subsequent studies, WGBSSuite (Rackham et al. 2015) could, for example, serve for a simulation of DMRs.

Furthermore, CPU-time, real-time, resident-set-size (rss) and virtual-set-size (vss) memory consumption should be variables to estimate the performance efficiency of the programs.

User friendliness, understood as the degree of installation and handling ease, has to be taken into account for project planning and the benchmarking of bioinformatics tools for the detection of differential methylation by WGBS.

3.7. Conclusions

We highlighted the current developments in the field of whole-genome bisulfite sequencing, the actual gold standard for 5mC analysis. The application of PCR-free libraries as well as direct methylation detection without bisulfite treatment through SMRT sequencing technologies is seen as a great advantage due to a lower PCR bias or the riddance of bisulfite treatment. However, due to high coverage recommendations and as a result of low base call accuracy, SMRT sequencing for large numbers of samples remains expensive at the moment. In future, this might be solved when lower error rates for these techniques are achieved.

3.8. Author Contributions

Conceptualization, C.G. and M.K.; methodology, C.G., M.K., C.B.; investigation, C.G., M.K., C.B.; writing—original draft preparation, C.G. and M.K.; writing—review and editing, M.K., C.B., B.G.; visualization, C.G.; supervision, B.G., I.G.; project administration, C.G., B.G., I.G.; funding acquisition, C.G., B.G., I.G.

3.9. Funding

This research was funded by the state of Saxony-Anhalt and Volkswagen Stiftung.

3.10. Conflicts of Interest

The authors declare no conflicts of interest.

3.11. Acknowledgments

We thank Samar Fatma, Alexander Gabel, Martin Posch and Marc Wagner for the valuable scientific discussions as well as Volkswagen Stiftung and the state of Saxony-Anhalt for their financial support.

4. Performance of Mapping Approaches for Whole-Genome Bisulfite Sequencing Data in Crop Plants

Published in Frontiers in Plant Science, 11, p. 176, 28 February 2020;
https://doi.org/10.3389/fpls.2020.00176

Special Issue: Advances in Applied Bioinformatics in Crops

Claudius Grehl [1,2*], Marc Wagner [3], Ioana Lemnian [1,4], Bruno Glaser [2], Ivo Grosse [1,5]

[1] Institute of Computer Science, Bioinformatics, Martin Luther University Halle–Wittenberg, Von Seckendorff-Platz 1, D-06120, Halle (Saale), Germany;

[2] Institute of Agricultural and Nutritional Sciences, Soil Biogeochemistry, Martin Luther University Halle–Wittenberg, Von–Seckendorff–Platz 3, D-06120, Halle (Saale), Germany;

[3] Freie Universität Berlin, Kaiserswerther Str. 16-18, D-14195, Berlin, Germany;

[4] Institute of Human Genetics, Martin Luther University Halle-Wittenberg, Magdeburger Str. 2, D-06112, Halle (Saale), Germany;

[5] German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Deutscher Platz 5e, D-04103, Leipzig, Germany

* Correspondence: claudius.grehl@informatik.uni-halle.de

4.1. Abstract

DNA methylation is involved in many different biological processes in the development and well-being of crop plants such as transposon activation, heterosis, environment-dependent transcriptome plasticity, aging, and many diseases. Whole-genome bisulfite sequencing is an excellent technology for detecting and quantifying DNA methylation patterns in a wide variety of species, but optimized data analysis pipelines exist only for a small number of species and are missing for many important crop plants. This is especially important as most existing benchmark studies have been performed on mammals with hardly any repetitive elements and without CHG and CHH methylation.

Pipelines for the analysis of whole-genome bisulfite sequencing data usually consists of four steps: read trimming, read mapping, quantification of methylation levels, and prediction of differentially methylated regions (DMRs). Here we focus on read mapping, which is challenging because un-methylated cytosines are transformed to uracil during bisulfite treatment and to thymine during the subsequent polymerase chain reaction and read mappers must be capable of dealing with this cytosine/thymine polymorphism.

Several read mappers have been developed for the last years with different strengths and weaknesses, but their performances have not been critically evaluated. Here, we compare 8 read mappers: Bismark, BismarkBwt2, BSMAP, BS-Seeker2, Bwameth, GEM3, Segemehl and GSNAP to assess the impact of the read-mapping results on the prediction of DMRs.

We use simulated data generated from the genomes of *Arabidopsis thaliana*, *Brassica napus*, *Glycine max*, *Solanum tuberosum*, and *Zea mays*, monitor the effects of the bisulfite conversion rate, the sequencing error rate, the maximum number of allowed mismatches, as well as the genome structure and size, and calculate precision, number of uniquely mapped reads, distribution of the mapped reads, run time, and memory consumption as features for benchmarking the eight read mappers mentioned above.

Furthermore, we validated our findings using real-world data of Glycine max and showed the influence of the mapping step on DMR calling in WGBS pipelines.

We find that the conversion rate has only a minor impact on the mapping quality and the number of uniquely mapped reads, whereas the error rate and the maximum number of allowed mismatches has a strong impact and leads to differences of the performance of the eight read mappers. In conclusion, we recommend BSMAP that needs the shortest run time and yields the highest precision and Bismark that requires the smallest amount of memory and yields precision and high numbers of uniquely mapped reads.

Keywords: epigenetics, DNA methylation patterns, read mapping, benchmarking, WGBS

4.2. Introduction

It has been shown that DNA methylation is involved in several biological mechanisms and diseases such as cancer (Koch et al. 2018). Plant methylation analysis is especially interesting as 5-methyl-cytosine (5mC) is involved in heterosis effect (Chen et al. 2018), transposon silencing and environment-dependent transcriptome plasticity (Lauss et al. 2018). However, beside the complementary CG methylation being highly abundant in animals, in plants CHG and uncomplimentary CHH (H = C, T or A) methylation have evolved from the former recognition system of foreign DNA.

Whole-genome bisulfite sequencing (WGBS) is often referred to as “gold standard” for 5mC detection because the whole genome is covered at a single-base resolution. Other methods cover only preselected genome regions enriched for cytosine-phosphate-guanine-dinucleotide (CpG) content or methylation, for example with the use of restriction enzymes in reduced representation bisulfite sequencing (rrBS) (Sun et al. 2015), or methylated DNA immune precipitation followed by next generation sequencing (MeDIP-seq) (Bock et al. 2010; Aberg et al. 2017).

Bisulfite-mediated conversion of unmethylated cytosines to uracil, and during PCR to thymine, leads to four different strands in the data sets after sequencing: original top, complementary to original top, original bottom and complementary to original bottom strand (Figure 2). Methylated cytosines remain unaffected and could be spotted by alignment of the generated sequencing reads to a reference genome or a non-bisulfite-treated control.

Critical within the bioinformatics analysis of WGBS data sets is the mapping step, as the reduced alphabet leads to specific challenges for the mapping tools due to the bisulfite treatment (Laird 2010).

In general, two different algorithmic approaches exist in bisulfite-read alignment tools for dealing with the unmethylated C to T conversion: the 'wild card' and the 'three letter' approach. In the wild card approach the Cs in the reference genome are replaced with the wild card 'Y' for pyrimidine bases and allows thus the alignment of Cs (methylated Cs) and Ts (possibly unmethylated Cs). The alignment itself is based on matching seeds (k-mers) to the reference and then extending them. In the three letter approach the alphabet of the genome and the reads is reduced to {A, G, T} by converting all Cs in the reference sequence and in the read data to Ts. Afterwards, the reads are mapped by conventional mappers such as Bowtie, Bowtie2 or bwa, so the alignment of bisulfite data profits directly by the improvements of traditional mappers.

One study already focussed on benchmarking of rrBS alignment using simulated rrBS and real human lung tissue data sets (Sun et al. 2018). Kunde-Ramamoorthy et al. (2014) evaluated the mapping performance of five mapping tools in WGBS datasets of human peripheral blood lymphocyte and hair follicle. Another study has been performed in plants, showing that the tool BismarkBwt2 performed best in terms of sensitivity and precision, not accounting for the coverage distribution across the reference genome (Omony et al. 2019). In contrast, our study is focussed on simulated WGBS in plants, covering different species with different amount of repetitive sequences. In addition, little is known about the mapping behaviour in crop plants. Furthermore, as all former studies did not systematically account for different parameter settings such as number of mismatches, we evaluate this parameter in more detail.

There is need for an extensive qualitative and quantitative benchmarking of alignment tools to avoid false interpretation of results in DNA methylation studies and to enable the application of precise, efficient and user-friendly pipelines. Especially a known “truth” is important, which could be generated by benchmarking datasets of simulated and, thus, known read data, for calculation of quality scores in multiclass hypothesis testing. In terms of quantitative comparison, time efficiency, amount of uniquely mapped reads and consumption of RAM has to be monitored, as well as the overall distribution of mapped reads to look for genomic regions with systematically lower coverages.

4.3. Material and Methods

Arabidopsis thaliana, *Brassica napus*, *Glycine max*, *Solanum tuberosum* and *Zea mays* (Table 1) have been examined to reveal potential inter-species variability in terms of mappability. These species have been chosen to cover different agronomically relevant plant families, different genome sizes and different assembly qualities. All reference genomes have been downloaded from http://plants.ensembl.org. We simulated WGBS datasets for 2 x 150 base pair paired-end reads for the five different plant genomes using the open-source WGBS simulation tool: Sherman (https://www.bioinformatics.babraham.ac.uk/projects/sherman/), which has been developed at the Babraham Institute. The reads have been simulated. We have chosen 150bp paired-end sequencing for our benchmarking study as it is the mostly applied and proposed sequencing option for WGBS experiments. Doing so, small repeats below the total fragment size of 500-700bp could be covered, which is especially important for repeat-rich (crop) plants. Furthermore, choosing a parameter set of 150bp paired-end facilitates the necessary multiplexing with non-bisulfite libraries during sequencing.

Table 1: Five species included in this benchmarking study with the size of the reference genome and the used reference genome version which has been taken for the simulation of the read datasets. The proportion of repetitive sequences is given as estimated value over the genome.

Species	*Arabidopsis thaliana*	*Brassica napus*	*Glycine max*	*Solanum tuberosum*	*Zea mays*
Genome Size (bp)	135,670,229	738,357,821	955,377,461	727,424,546	2,104,350,183
Genome version	TAIR10	AST_PRJEB5043_v1	Glycine_max_v2.1	SolTub_3.0	B73 RefGen_v4
Repeats in %	<23 (Flutre et al. 2011)	~48 (Liu et al. 2018)	~57 (Schmutz et al. 2010)	~49 (Mehra et al. 2015)	~75 (Wolf et al. 2015)
Citation	Lamesch et al. 2012	Chalhoub et al. 2014	Schmutz et al. 2010	Xu et al. 2011	Schnable et al. 2009

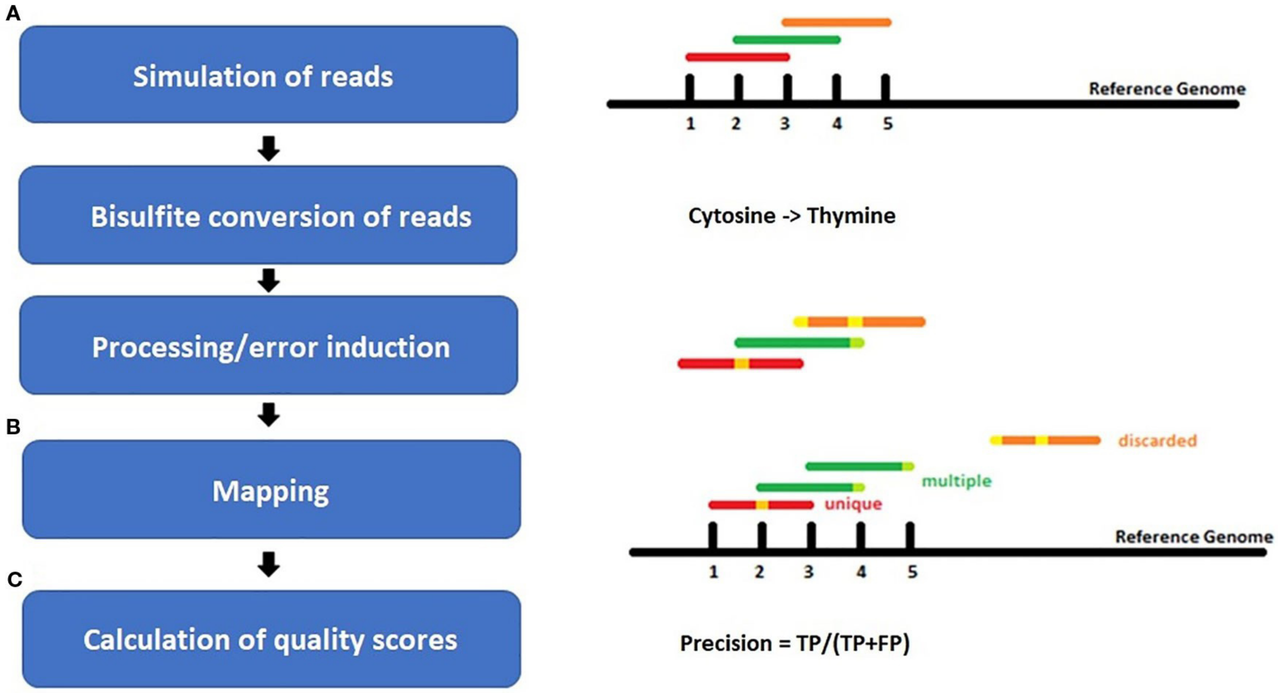


Figure 6: Workflow of the experiment setup for (A) simulation of bisulfite-treated reads based on a reference genome using the tool Sherman, including bisulfite conversion and error induction. Afterwards (B) mapping of the simulated datasets and (C) calculation of quality scores. The color coding shows the different classes of reads after mapping: red = uniquely mapped, green multiply mapped, orange= discarded/unmapped reads.

For each species, benchmarking datasets in 5-fold sequencing depth, three bisulfite conversion rates [90 %, 98 %, 100 %] and four different sequencing error rates [0 %, 0.1 %, 0.5 %, 1 %] have been simulated. The sequencing errors have been modelled to account for more likely errors at the end of a read, like in real world sequencing data (Figure 6). Whereas the overall resulting phred score of 30 is equivalent to an error rate of 0.1 % or 1 in 1000 wrong base calls. Illumina HiSeq sequencing yields an error rate of 0.0034-1 % while PacBio shows 5–10 % false base calls (Escalona et al. 2016). We decided to include 98 % conversion rate as this is usually guaranteed by sequencing facilities and 90 % to look for a value below this threshold.

For mapping the simulated WGBS reads to the genomes, we tested several wild-card and three-letter mappers: Bismark (Krueger and Andrews 2011), BSMAP (Xi and Li 2009), BS-Seeker2 (Guo et al. 2013), Bwa-meth (Pedersen et al. 2014), GEM3 (Marco-Sola et al. 2012), GSNAP (Wu and Nacu 2010), and Segemehl (Hoffmann et al. 2009; Otto et al. 2012). These mappers differ in term of their "age", number of citations and indexing strategy (Table 2). For further insight into mapping and indexing strategies as well as for an insight into the underlying algorithmic approaches we could recommend (Tran et al. 2014).

Table 2: Bisulfite Read mapping tools evaluated in this survey, listed by their mapping and indexing strategy. The number of citations is based on the Web of Science Core Collection (date: 19.1.2020).

Mapper name	Strategy	Indexing	Version	Citations
Bismark	3 letter	BWT (bowtie 2/bowtie 1)	0.19.1	1.176
BSMAP	wild-card	Hash table (SOAP)	2.73	3
BS-Seeker2	3 letter	BWT (bowtie 2)	2.1.5	135
Bwa-meth	3 letter	BWA mem	0.2.2	3
GEM3	3 letter	Custom FM index	3.6.1-2	236
GSNAP	wild-card	Hash table	2019-06-10	83
Segemehl	wild-card	Enhanced suffix array	0.2.0	283

Bismark (Krueger and Andrews 2011), one of the most cited 3 letter mapper for bisulfite-sequencing data, first converts the reads and the genome into two versions: a C-to-T and in a G-to-A versions. Afterwards the two read versions are aligned to the two versions of the reference genome with the goal of detecting to which of the fours strands (Figure 2) the read fits. This alignment is performed by four parallel instances of either Bowtie (Langmead et al. 2009), one of the fastest mapper for NGS data, or Bowtie2 (Langmead and Salzberg 2012), an improved version of bowtie, that allows gapped alignment.

BSMAP (Xi and Li 2009) is included in the list for being the first mapper for the alignment of bisulfite data. It uses an efficient HASH table seeding algorithm for indexing the genome, bitwise masking each nucleotide in the reads and the reference and matching them each other in an efficient way.

GSNAP (Wu and Nacu 2010) is a general purpose mapper that can also deal with bisulfite data. Like BSMAP, it is based on special hash tables and uses a wild-card approach to match read seeds to genofme regions. Since its original publication several improvements of the algorithms have been made by increasing the length of the hashed k-mers, adding a compression mechanism and using enhanced suffix arrays (Wu et al. 2016).

BS-Seeker2 (Guo et al. 2013) is the extension of BS Seeker (Chen et al. 2010) for mapping bisulfite data and deploys a three letter approach. In addition to performing a gapped alignment it is able to filter out reads with incomplete bisulfite conversion, in this way increasing the specificity.

Compared to the other tools in this benchmark Bwa-meth (Pedersen et al. 2014) is a relatively new mapper for bisulfite data. It is based on BWA-mem aligner (Li and Durbin 2009, 2010) and it is advertised as a fast and accurate aligner.

Segemehl was originally designed as a general purpose mapper (Hoffmann et al. 2009) but has been extended to handle bisulfite data (Otto et al. 2012). Segemehl achieves a high sensitivity by using a wild-card approach based on enhanced suffix arrays for the seed search and the Myers bit-vector algorithm for computing semi-global alignments.

The 8 mappers have been used for mapping of the simulated reads with 0, 1, 2 and 3 mismatches in the read allowed. As Bowtie2 and Bwa-meth do not allow setting the total number of mismatches in a read as parameter, but in the seed, we performed our analysis on the basis of 0 mismatches in the seed for these mapping tools. Other parameter settings, such as number of threads used, have been the default values of the mentioned tools if not stated otherwise and are comparable across the different tools. All scripts are available at git-hub (https://github.com/grehl/benchWGBSmap).

After mapping, the reads can be classified in three different classes: i) discarded reads that could not be mapped, ii) multiple mapped reads that could be aligned but to more than one position on the reference genome because of sequence similarities, and iii) uniquely mapped reads, which have been mapped to one position only.

For further evaluation, we used only uniquely mapped reads. Since we did not account for insertions and deletions, we have considered only the first base of the read at its genome position. When calculating the quality scores, we have compared the true and the predicted position of a read. For each read the true genome origin is known, since Sherman encodes it in the read name, while the predicted position is derived from the alignment files.

The quality scores computed are the amount of considered unique reads, the precision, the memory consumption and the time consumption of the tools. Furthermore, we looked at the read distribution over the reference genome to account for systematic mapping deficiencies.

The precision of a mapping tool for a data set has been computed by using the formula for macro-averaged precision (macroAvgPrecision; TP = true positives, FP = false positives):

$$macroAvgPrecision = \frac{\sum_{i=1}^{N} \left(\frac{TP_i}{TP_i + FP_i}\right)}{N}$$

We first calculated the precision for every position i, summed over all positions and divided by the total number of positions N. The macro-averaging has been chosen as it weights FP higher than in micro-averaging calculation of the precision. We furthermore use “precision” in this chapter instead of “macro-averaged precision”.

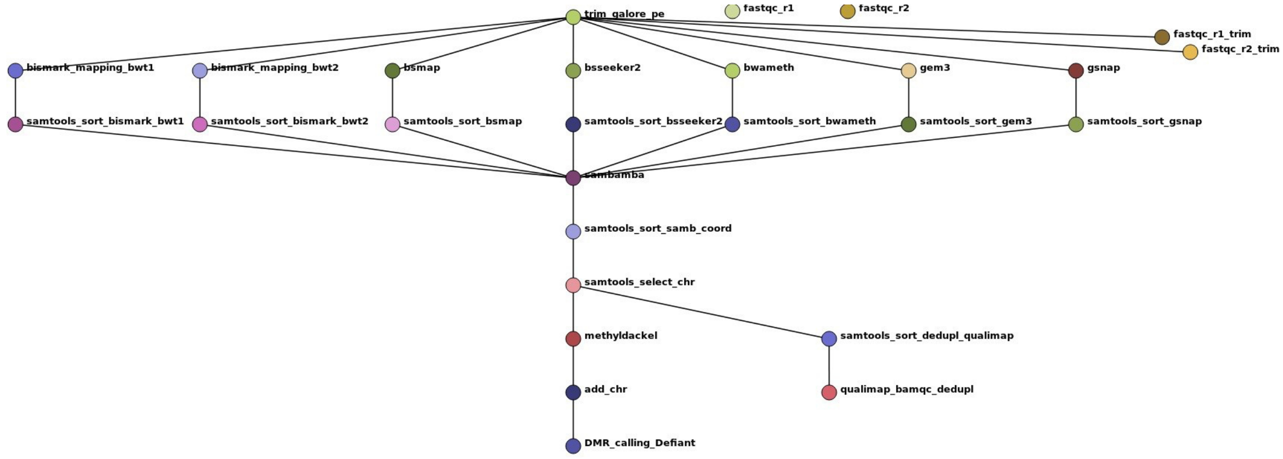


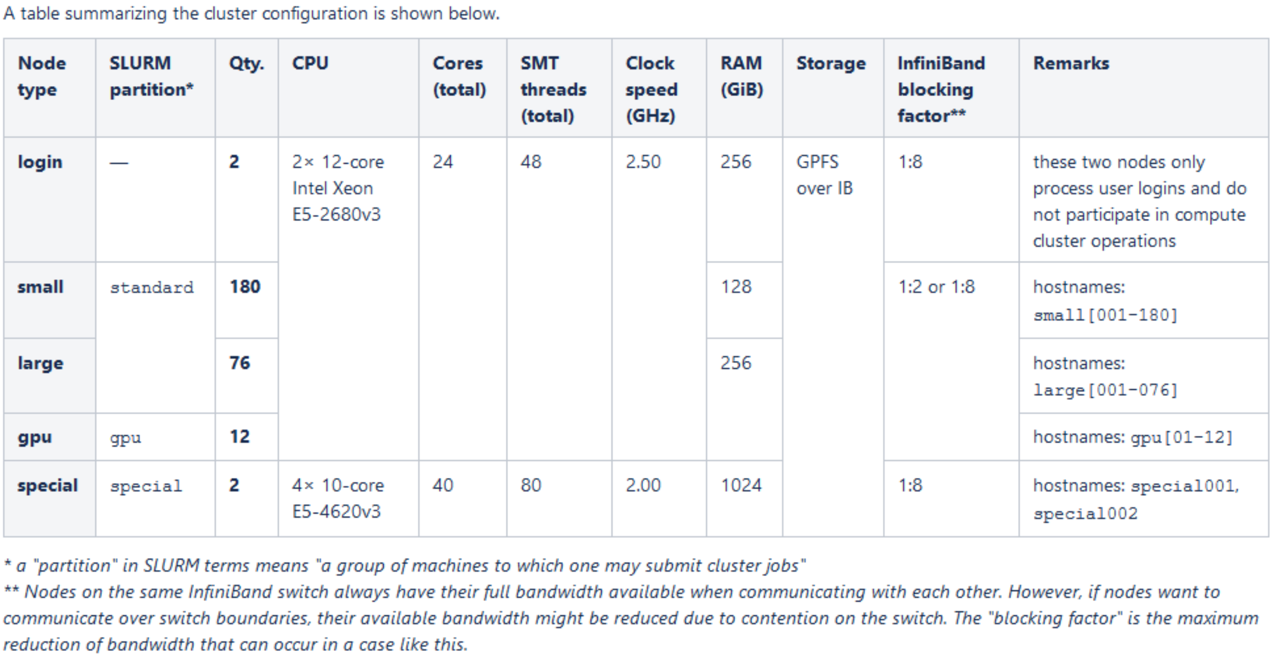
Figure 7: Workflow show as Directed Acyclic Graph (DAG) of the benchmark experiments using 7 mapping tools on a real-world dataset of *Glycine max* root hair grown under two different temperatures to exemplarily compare the mapping performance and the influence of the mapping tools on DMR calling. The workflow includes two quality checks on the raw reads prior and after trimming with trim galore using fastqc, the mapping, a samtools sorting step to account for different output formats of the tools, a sambamba deduplication step, coordinate sorting, exclusion of scaffold mappings, methylation level calling using methyldackel and after adding “chr” to the methylation report output the final DMR calling with Defiant. To evaluate the bam quality we used the qualimap bamqc function.

To evaluate the impact of the tested tools on DMR calling and to show the reliability of our simulated benchmark study, we included a real-world dataset of *Glycine max* root hair samples grown under 25 °C and 40 °C (Hossain et al. 2017). For automatization we implemented a snakemake pipeline (Köster and Rahmann 2012), which is shown in Figure 7. Other tools used in this pipeline are: trim galore (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/), fastqc (Leggett et al.

2013), qualimap (García-Alcalde et al. 2012), samtools (Li et al. 2009), sambamba (Tarasov et al. 2015), methyldackel (https://github.com/dpryan79/MethylDackel), Defiant (v.1.1.6) (Condon et al. 2018) [parameter settings: -c 10 -v 'BY' -CpN 5 -p 0.05 -P 10] and Circos (Krzywinski et al. 2009). The mapping parameter sets were comparable and allowed 0 mismatches. As Segemehl showed an extensive memory consumption and runtime we had to exlude this mapper for the qualitative benchmark study and the mapping with the real dataset. For the DMR calling we had to exclude BS-Seeker2 as the flag information did not follow standard formats so the files could not reliably be used for methylation calling. The settings for DMR calling have been: minimum 10fold coverage, minimum 5 CpN in one DMR, minimum +/-10 % methylation difference between the two treatments with a maximum p-value of 0.05. Analogous to the source code of the simulated benchmark study, we mainly relied on the default parameters if not stated otherwise in the script. All scripts are available online at git-hub (https://github.com/grehl/benchWGBSmap).

Simulation, mapping and quality score calculation has been performed on IANVS High-Performance-Cluster of Martin-Luther University Halle-Wittenberg (Table 3). For calculation of runtime and memory consumption only one core of the login-nodes has been allowed for mapping. For simplicity, to give an overview about mapping, and as the genome size is the most important factor with respect to runtime and memory consumption, we decided to focus on this factor only. A subsequent study could focus on the influence of other factors of runtime and memory consumption. The mapping of the quality benchmark has been performed on the "large" and "small" nodes.

Table 3: IANVS Cluster Specifications.



A table summarizing the cluster configuration is shown below.

Node type	SLURM partition*	Qty.	CPU	Cores (total)	SMT threads (total)	Clock speed (GHz)	RAM (GiB)	Storage	InfiniBand blocking factor**	Remarks
login	—	2	2× 12-core Intel Xeon E5-2680v3	24	48	2.50	256	GPFS over IB	1:8	these two nodes only process user logins and do not participate in compute cluster operations
small	standard	180					128		1:2 or 1:8	hostnames: small[001-180]
large		76					256			hostnames: large[001-076]
gpu	gpu	12								hostnames: gpu[01-12]
special	special	2	4× 10-core E5-4620v3	40	80	2.00	1024		1:8	hostnames: special001, special002

** a "partition" in SLURM terms means "a group of machines to which one may submit cluster jobs"*
*** Nodes on the same InfiniBand switch always have their full bandwidth available when communicating with each other. However, if nodes want to communicate over switch boundaries, their available bandwidth might be reduced due to contention on the switch. The "blocking factor" is the maximum reduction of bandwidth that can occur in a case like this.*

4.4. Results

The results of our quantitative benchmark studies for memory consumption (Figure 8) and runtime (Figure 9) are shown for the 8 mappers in relation to the size of the reference genome. The memory consumption ranged from 0.1 GB for the mapping of the *Arabidopsis thaliana* dataset with Bismark either using bowtie or bowtie2 to 39 GB for the mapping of the *Zea mays* dataset with Segemehl. All datasets had 100 % conversion rate, 0 % error rate and 0 mismatches allowed during the mapping. Similar patterns in the memory consumption and runtime have been observed also for datasets with other parameter settings. In terms of runtime, the user time is depicted, ranging from few minutes for all mappers using the *Arabidopsis thaliana* dataset, to 79 h for the mapping of a *Zea mays* dataset with Segemehl. Overall BSMAP took the least time, especially for large reference genomes. It is interesting to note that althought *Solanum tuberosum* and *Brassica napus* have a similar genome size, some mappers had a higher memory consumption (Segemehl, GEM3, GSNAP) and runtime (Segemehl, BismarkBwt2, GSNAP) for *Brassica napus*. This might be due to the large amount of large repetitive regions and paralogue genes within the *Brassica napus* genome, as the overall proportion of repeats is comparable to *Solanum tuberosum*.

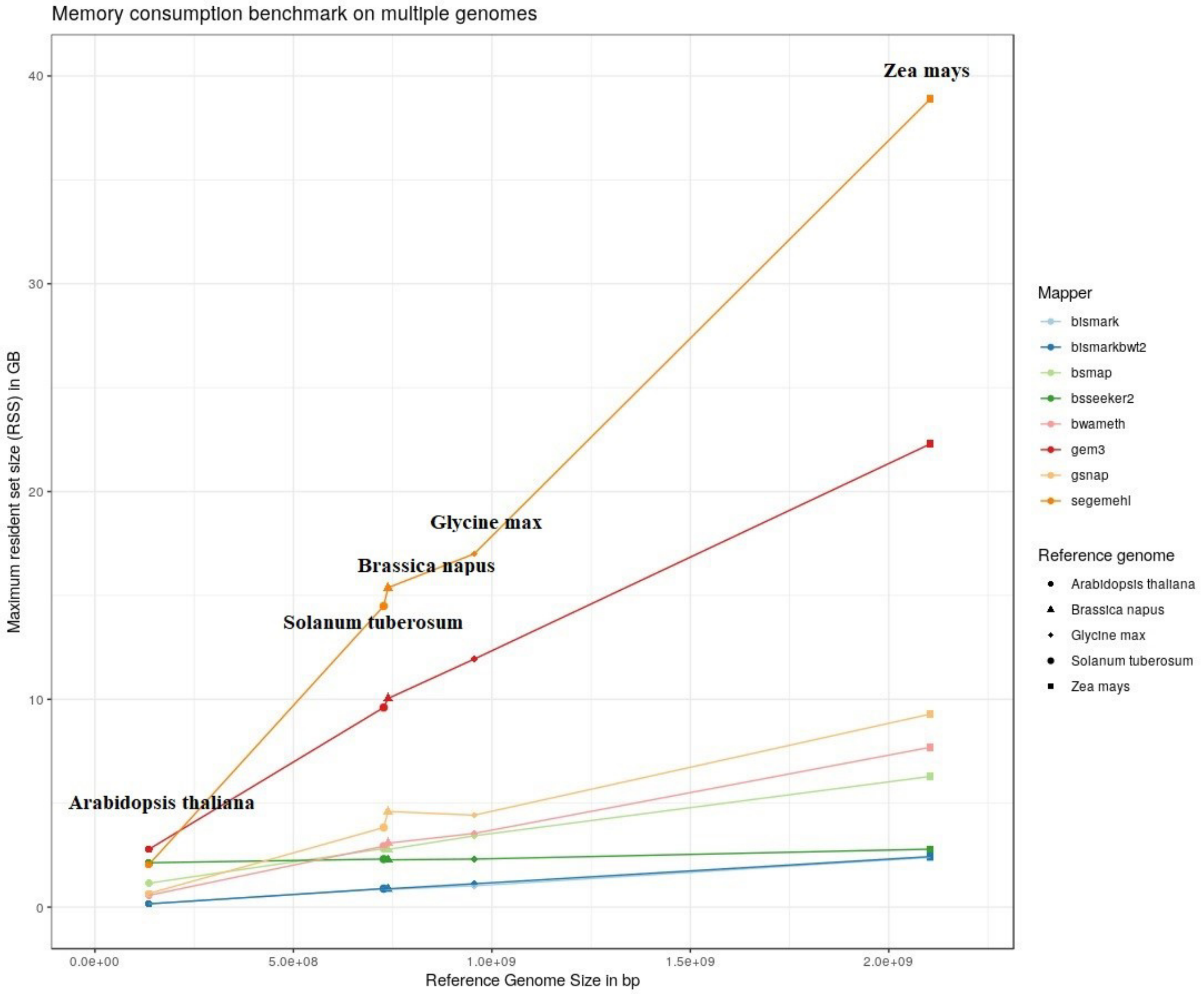


Figure 8: Maximum resident set size in GB of 8 mappers for 5fold simulated bisulfite converted datasets out of five reference genomes (*Arabidopsis thaliana, Brassica napus, Glycine max, Solanum tuberosum, Zea mays*). 0 % error rate, 100 % conversion rate, 0 mismatch allowed.

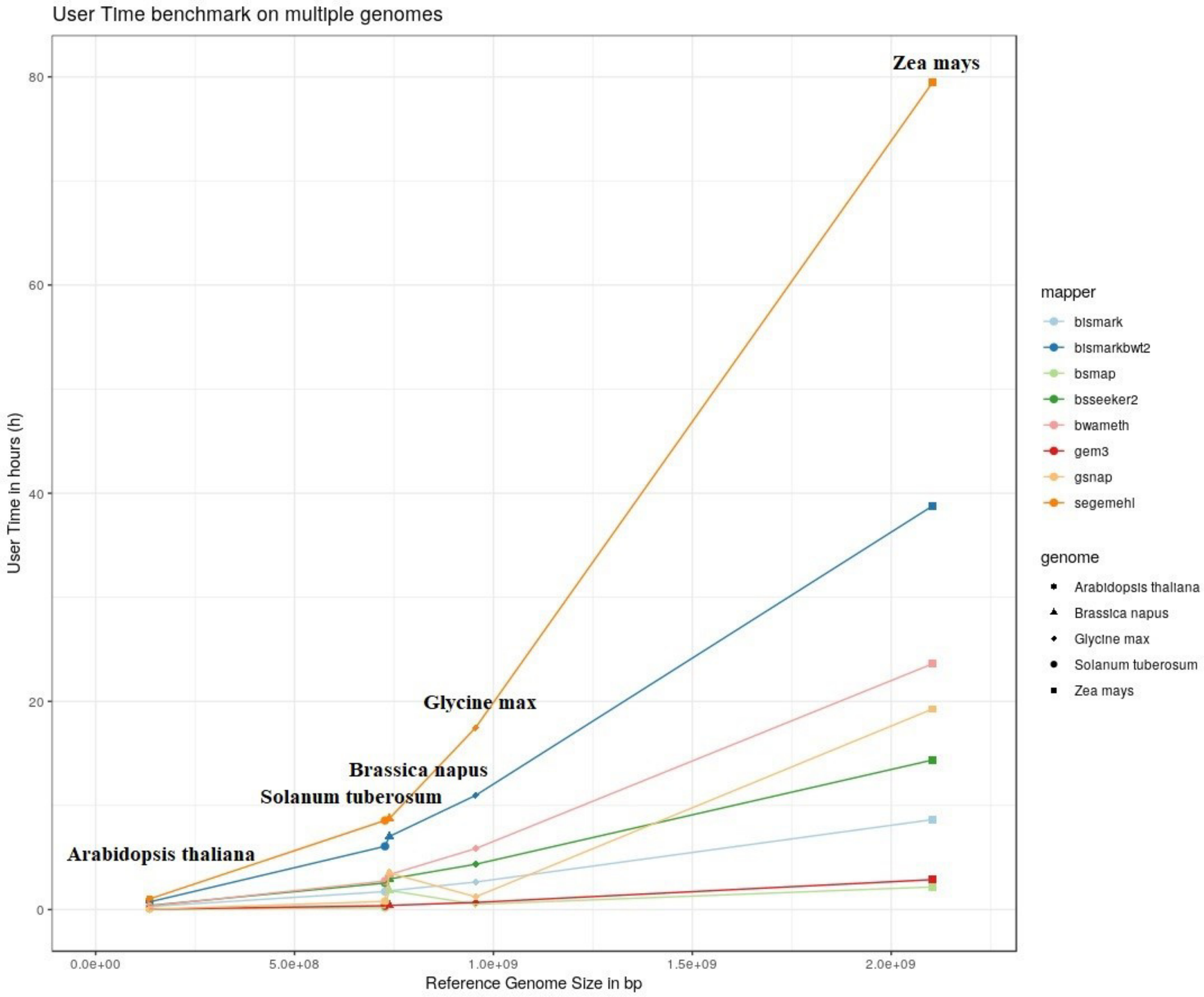


Figure 9: User timer in hours of 8 mappers for 5 fold simulated bisulfite converted datasets out of five reference genomes (*Arabidopsis thaliana, Brassica napus, Glycine max, Solanum tuberosum, Zea mays*). 0 % error rate, 100 % conversion rate, 0 mismatch allowed.

Because of the extensive memory consumption and runtime of Segemehl, we excluded this mapper from the quality benchmark study.

Overall, the conversion rate did not influence the number of uniquely mapped reads or the mapping quality (supplementary material). In terms of the mapping quality, in relation to the error rate, and the reference genome, we basically see three groups of mappers (Figure 10). The first group is independent of the allowed number of mismatches during the mapping and includes Bismark, BismarkBwt2, Bwa-meth and GEM3. The second group consists of BSMAP and BS-Seeker2, showing an increase in the number of uniquely mapped reads with higher numbers of allowed mismatches with barely no changes in precision. The third group including GSNAP shows an increase in the number of uniquely mapped reads but a decrease in the precision with higher numbers of mismatches allowed. As BismarkBwt2 and bwameth do not allow to set the number of mismatches in the entire read, both are labelled with a triangle. Between the analysed genomes we see differences for all mappers with the tendency to lower numbers of uniquely mapped reads in *Zea mays* and lower precision in *Zea mays* and *Brassica napus*.

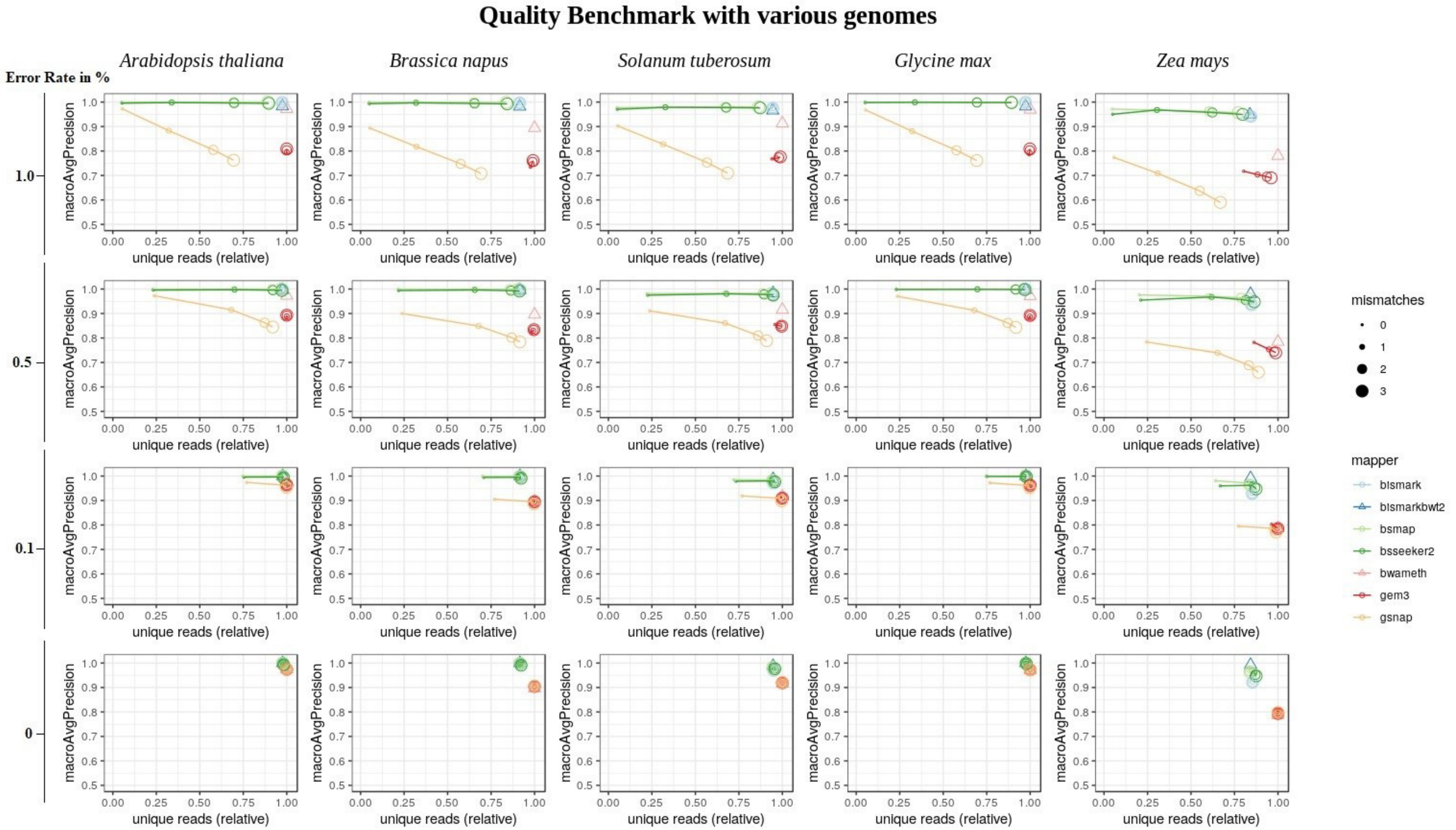


Figure 10: Quality benchmark of 7 mappers based on simulated bisulfite sequencing datasets in *Arabidopsis thaliana*, *Brassica napus*, *Glycine max*, *Solanum tuberosum*, and *Zea mays*. We simulated the datasets with 4 different error rates [0, 0.1, 0.5 and 1 %] in a 5fold coverage. For 5 out of 7 mappers we had the opportunity to allow for different numbers of mismatches [0, 1, 2, 3]. These mappers are depicted by circles. Two mappers, bismark using bowtie2 and bwameth, did not allow the adjustment for different numbers of mismatches in the entire read. They are depicted by triangles. The conversion rate had no effect and is therefore not shown in this figure. The depicted conversion rate is 100 % for all data sets.

For *Arabidopsis thaliana* (Figure 11) and *Glycine max* (Figure 12) the distribution of reads over the reference genome is exemplarily shown for the mapping of two datasets each. The first dataset has been simulated with 100 % conversion rate, 0 % error rate and has been mapped with 0 mismatches allowed for all 7 mapping tools, depicted in the lower window. The upper window shows 100 % conversion rate, 1 % error rate and 0 mismatches, again for all 7 mapping tools. All coverage plots have a resolution of 400 windows across the whole reference genome. For higher error rates BSMAP, BS-Seeker2 and GSNAP show a severe decrease in coverage. Furthermore, we clearly see several regions with a decrease in coverage within the reference genome independent of the error rate. In grey, we highlighted the regions which are known to contain a high percentage of repetitive sequences. Bismark and BismarkBwt2 are depicted behind each other, showing nearly the same coverage distribution. In total, Bwa-meth shows the least derivation in the coverage distribution.

The benchmarking of the real *Glycine max* dataset resulted in proper mapped paired-end read counts leading to the mean coverage shown in Table 4. The last column shows the final number of DMRs. These are additionally depicted in Figure 13.

Table 4: Mean coverage of the 4 real data samples and the results of the DMR calling (SRR5044695 & SRR5044696 are the control and SRR5044699 & SRR5044700 are the heat stress replicates).

	SRR5044695	SRR5044696	SRR5044699	SRR5044700	DMRs
BismarkBwt2	15,0	14,5	17,9	19,2	281
Bwa-meth	32,5	29,5	35,6	38,8	256
GEM3	28,3	25,8	31,4	34,3	136
BismarkBwt1	11,3	11,0	13,9	14,7	97
GSNAP	10,3	9,5	12,0	12,6	70
BSMAP	10,5	9,9	12,2	12,7	63
BS-Seeker2	8,6	8,4	11,0	11,3	X

4.5. Discussion

We performed an extensive benchmarking experiment based on simulated data to evaluate the qualitative and quantitative performance of mappers for bisulfite sequencing data in 5 plant species with focus on crop plants.

In terms of user time and memory consumption, the different tools showed big differences. Especially for larger genomes, for example Segemehl used a tremendous amount of RAM and needed the most time to map the given reads onto the reference genome. For larger reference genomes (>4 GB) the genome has to be splitted if Segemehl should be used. For these two reasons, we could not use Segemehl for mapping of huge datasets such as *Zea mays*, even as it performed well in terms of precision in a pilot study. BSMAP, GEM3 and GSNAP showed only a low increase in time with increasing size of the genome, but used more memory. Especially Bismark showed a low increase for the memory consumption and a relatively low increase in run time. The high difference between Bismark and BismarkBwt2 comes most likely due to the “soft clipping” function of BismarkBwt2.

The mapping quality and number of uniquely mapped reads changes between the tools with *Zea mays* showing the lowest precision scores and the least number of uniquely mapped reads. This effect might be caused by the high amount of repetitive sequences, which has been shown to make up to 75 % of the *Zea mays* genome containing mostly gypsy- and copia-like long, terminal repeats (LTR) (Wolf et al. 2015). For *Glycine max* the described amount of repeats lays around ~57 %. This also includes telomeric as well as centromeric repeats and not annotated repeats where the reference genome shows scaffolded regions (Schmutz et al. 2010). A wild-type reference genome sequencing consortium recently found 54 % repeats (Xie et al. 2019). As most repeats are <50 bp (Sherman-Broyles et al. 2014), the 2 x 150 bp paired-end reads with an insert size of 200 bp – 400 bp could cover large parts of the genome uniquely. The distribution of reads across the reference genome shows a good overlap with known and long, repeat-rich regions. Some mappers such as GEM3 and GSNAP tend to map high amounts of FP in this regions. Other mappers leave these regions out, leading to a lower coverage.

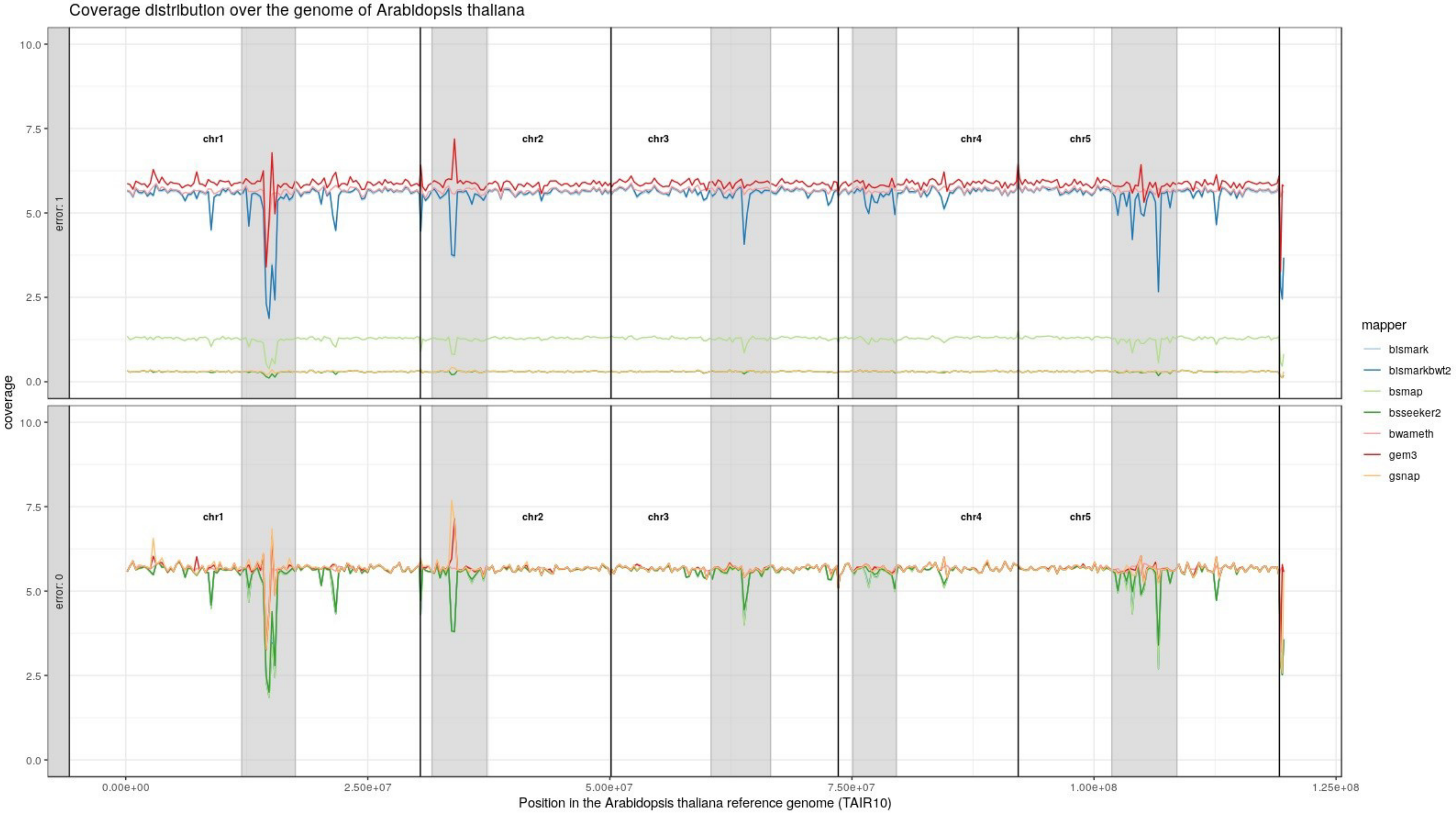


Figure 11: Coverage distribution over the reference genome of *Arabidopsis thaliana* (TAIR10). The lower window shows the performance of 7 mapping tools using a simulated 5fold coverage dataset with 0 % induced error rate, 100 % conversion rate and 0 mismatches allowed. The upper window shows a simulated 5fold coverage dataset with an induced error rate of 1 %, 100 % conversion rate and with 0 mismatches allowed during the mapping. The number of reads has been calculated based on the ensemblPlants “Base Pairs” information. This could cause small differences to the estimated 5fold coverage datasets. Black lines indicate the borders of chromosomes. Grey regions highlight highly repetitive regions.

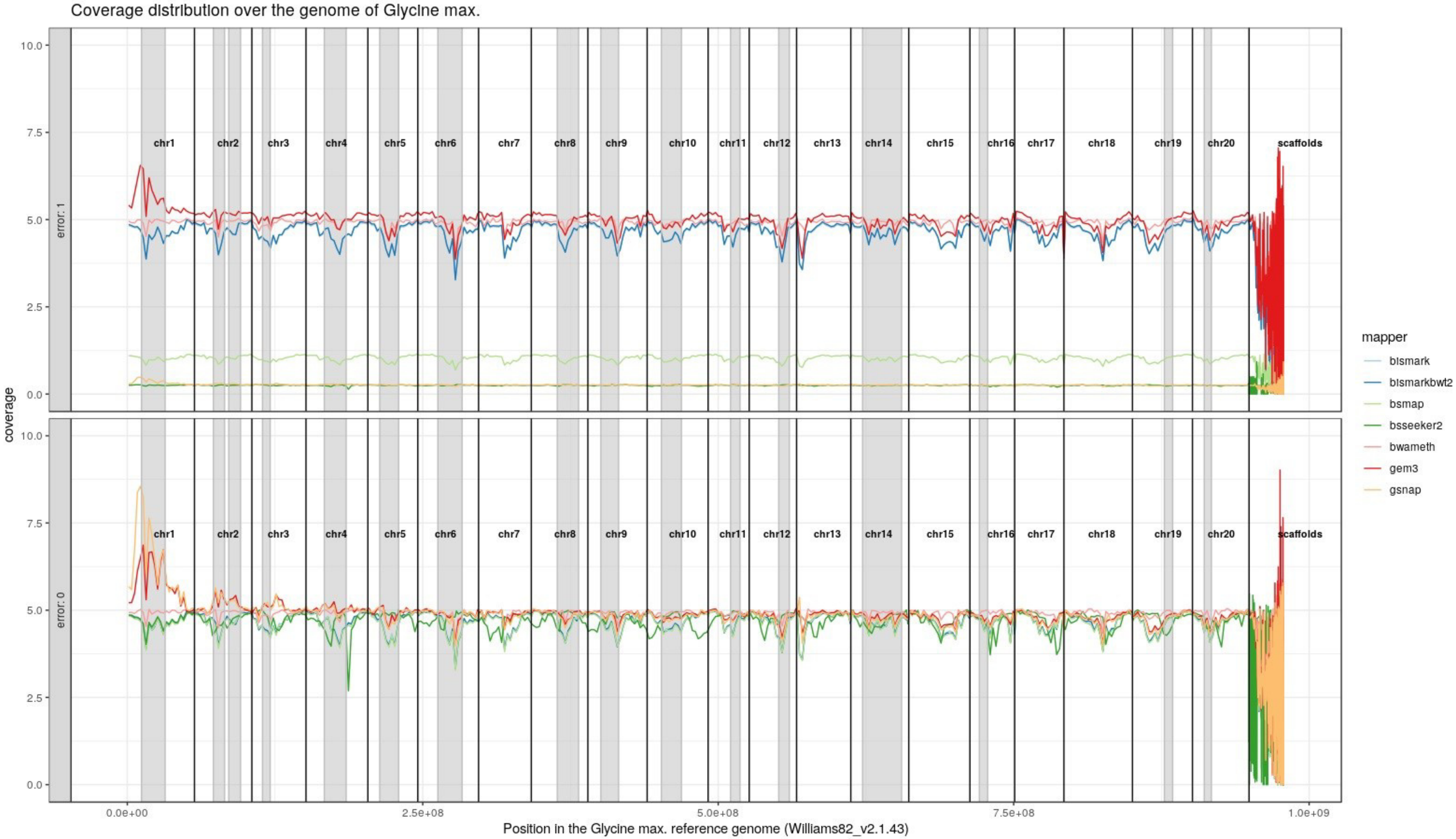


Figure 12: Coverage distribution over the reference genome of *Glycine max* (Williams82_v2.1.43). The lower window shows the performance of 7 mappers using a simulated 5fold coverage dataset with 0 % induced error rate, 100 % conversion rate, and 0 mismatches allowed. The upper window shows a simulated 5fold coverage dataset with an induced error rate of 1 %, 100 % conversion rate, and with 0 mismatches allowed during the mapping. Black lines indicate the borders of chromosomes. Grey regions highlight highly repetitive regions

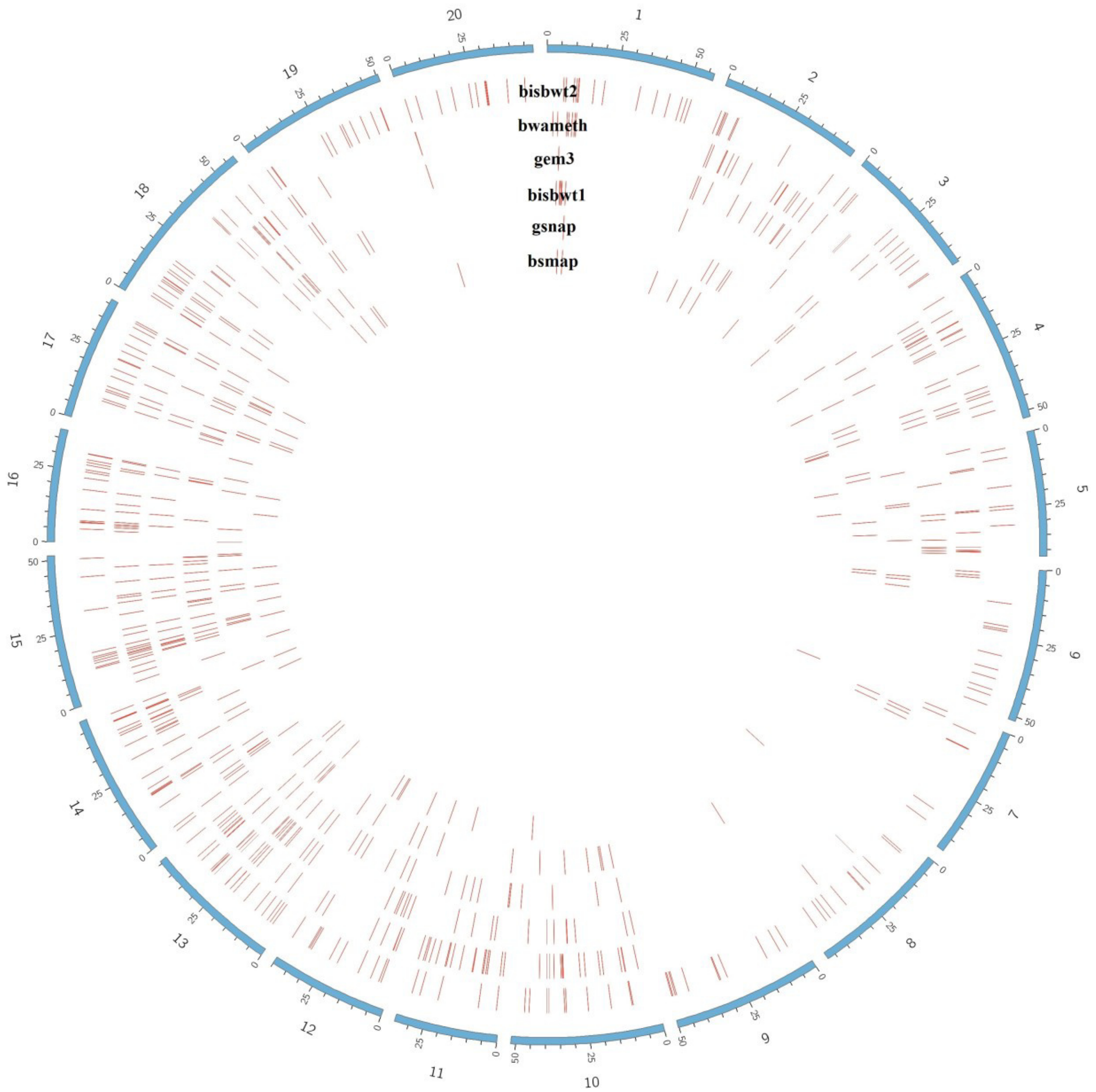


Figure 13: Circular plot showing the distribution of DMRs (red) as result of mapping the same dataset with 6 different mapping tools on the Glycine max_v2.1 reference genome. The outer circle shows the chromosomes of *Glycine max* in blue. Blue lines indicate hypomethylation, whereas red lines indicate hypermethylation (see the full list of DMRs at https://github.com/grehl/benchWGBSmap). Numbers of overlapping DMRs could be found in the Supplementary Material.

For the second part of our study we mapped the same datasets with the 7 mentioned mapping tools, but had to exclude BS-Seeker2 for the DMR calling. Here, we see the most unique, proper paired reads for mapping with Bwa-meth and GEM3. Surprisingly, this could not be confirmed for the DMR calling where we got the most DMRs using Bismark with Bowtie2 under usage of the same parameter sets, the same tools and the same pipeline. We could only speculate what the reason for this behaviour might be. Most likely this shift in the performance difference between the tools could be caused by false positive mappings which did not heavily influence the DMR calling as they might have been mapped to "non-sense" positions either already involved in a DMR region, not causing much harm in remote regions due to the coverage threshold of 10fold or they have been evenly distributed over treatment and control datasets.

In terms of precision, runtime and power to detect CpG-sites, (Sun et al. 2018) found Bwa-meth and BS-Seeker2 to be the best tools based on simulated and real rrBS reads from human lung tumor tissue. However, this stands in contrast to our findings, which show precision deficiencies for Bwa-meth under error rates above 0.1 % especially in repeat-rich and large plant genomes. BS-Seeker2 mapped reads precisely but error rates above 0.5 % and 0 mismatches allowed during mapping leads to unique mapping rates below 25 %. Other studies found Bismark to yield a reasonable combination of low memory consumption, low runtime and high quality scores (Kunde-Ramamoorthy et al. 2014; Omony et al. 2019). This could be confirmed by our study, where Bismark showed the lowest memory consumption in all tested genomes. For runtime, we see high differences between Bismark using Bowtie and BismarkBwt2 under usage of Bowtie2. The precision showed good scores for all genomes and settings, with a slight decrease for the *Zea mays* genome.

In conclusion, we have shown high differences between the available mapping tools for bisulfite-treated reads based on simulated and real datasets in terms of runtime, memory consumption and mapping quality. We see the stability of mapping quality against changes in the conversion rate, high differences between the mapping tools in terms of the number of uniquely mapped reads as well as in the capability to map correctly under the impact of higher error rates in 5 different genomes. Additionally, we see high differences with regard to the analysed genome, dependent on the size and structure of repeats.

For *Arabidopsis thaliana* basically every of the examined mapping tools could be used with a sufficient mapping rate and good quality, at least when assuming a low error rate. This holds true for low error rates in *Glycine max* mappings. For higher error rates we recommend Bwa-meth aswell as Bismark using Bowtie1 or Bowtie2. For paralogue-rich species such as *Brassica napus*, polyploid species such as *Solanum tuberosum* or large genomes with many repetitive sequences such as in *Zea mays* we prefer correct mappings over a large number of unique mapped reads. Therefore going with Bismark using Bowtie1 or Bowtie2 or BSMAP and BS-Seeker2 with a higher number of mismatches allowed might work well out of the perspective of mapping.

All in all, we recommend BSMAP that needs the shortest run time and yields the highest precision and Bismark that requires the smallest amount of memory and yields high precision and high numbers of uniquely mapped reads. Furthermore, Bwa-meth could be used with care in terms of precise calling of DMRs.

4.6. Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation, to any qualified researcher.

4.7. Author contributions

Conceptualization: CG and MW Methodology: CG, MW Investigation: CG, MW. Writing—original draft preparation: CG and IL. Writing—review and editing: IL, BG, IG. Visualization: CG and MW. Supervision: BG, IG. Project administration: CG, BG, IG. Funding acquisition: CG, BG, IG.

4.8. Funding

This research was funded by the state of Saxony-Anhalt and Volkswagen Stiftung. The Article Processing Charges have been supported by the open-science fund of the Martin-Luther University Halle-Wittenberg. IL was funded by the Deutsche Forschungsgemeinschaft (DFG, Germany) Research Training Group 2155 (ProMoAge).

4.9. Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

4.10. Acknowledgments

We thank all the members of the Bioinformatics Group of Halle-Wittenberg University, especially Alexander Gabel, Jan Grau, Claus Weinholdt, Silvio Weging, and Samar Fatma for valuable discussions as well as for support in terms of biostatistics and visualization.

4.11. Supplementary Material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2020.00176/full#supplementary-material

5. Differences in DNA methylation patterns of organically and conventionally grown potato (*Solanum tuberosum*) samples

Manuscript, submission planned

Claudius Grehl [1, 6*], Markus Kuhlmann [2], Paul Mäder [3], Jochen Mayer [4], Lothar Altschmied [2], Ivo Grosse [1, 5], Bruno Glaser [6]

[1] Martin Luther University Halle-Wittenberg, Institute of Computer Science, Von Seckendorff-Platz 1, 06120 Halle (Saale), Germany; claudius.grehl@informatik.uni-halle.de

[2] Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Corrensstraße 3, 06466 Gatersleben, Germany; kuhlmann@ipk-gatersleben.de

[3] Research Institute of Organic Agriculture, Ackerstrasse 113, CH-5070 Frick, Switzerland; paul.maeder@fibl.org

[4] Agroscope, Forschungsbereich Agrarökologie und Umwelt, Reckenholzstrasse 191, CH-8046 Zürich, Switzerland; jochen.mayer@agroscope.admin.ch

[5] German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Deutscher Platz 5e, 04103 Leipzig, Germany; ivo.grosse@informatik.uni-halle.de

[6] Martin Luther University Halle-Wittenberg, Institute of Agricultural and Nutritional Sciences, Soil Biogeochemistry, von-Seckendorff-Platz 3, 06120 Halle (Saale), Germany; bruno.glaser@landw.uni-halle.de

* Correspondence: claudius.grehl@informatik.uni-halle.de

5.1. Abstract

DNA methylation is a heritable epigenetic mark. For instance, DNA methylation patterns can be specific to environmental conditions. In our approach, differences in DNA methylation pattern were analysed in tubers of potato (Solanum tuberosum) that have been grown under organic and conventional farming conditions. These conditions differed in application of fertilizer, herbicides, fungicides and insecticides. Post-bisulfite-adapter tagging whole-genome bisulfite-sequencing (PBAT-WGBS) was performed on samples from two different years under organic and conventional growing conditions in three independent field replicates to identify differentially methylated regions. One of the identified clusters was associated with the StATOX1 gene. StATOX1 is an evolutionarily highly conserved antioxidant copper chaperone, responsible for detoxification of copper in organisms. Copper content of the potato samples was measured by Inductively-Coupled Plasma-Atomic-Emission-Spectrometry (ICP-AES) after high pressure nitric acid oxidation. As organic potato management relies on the application of copper as fungicide, our identified region might be indicative for the application of this specific compound.

Keywords: Plant epigenetics, Solanum tuberosum, DNA methylation, whole-genome bisulfite sequencing, organic agriculture, product authentication, food fraud.

5.2. Introduction

Organic farming aims at a sustainable use of natural resources, bio-based circular economy, a higher biodiversity and an environmentally friendly way to produce food and other products such as clothes and cosmetics (Mader et al. 2002). Up to now, no global standard is defined for organic farming. In the European Union (EU), rules for production, labeling and control are defined by European Commission regulation No 889/2008 (European Commission 2008). Within the EU, for growth of organic products, no fertilization with synthetically produced nitrogen sources is allowed (Haber-Bosch process). Nitrogen supply in organic farming is based on a high percentage of legumes in the crop rotation and the reuse of symbiotically fixed nitrogen by manure, compost and other organic sources. For weed, pest and disease control, plant protection takes place based on diverse crop rotations, mechanical weeding and biological pest control with antagonists and pesticides obtained from mostly natural sources such as copper. Further, application of genetically modified material for feeding and breeding is excluded. For potato production, the farmers are restricted to the use of copper ions to prevent potato late blight caused by *Phytophtora infestans*. Currently, the authentication is guaranteed through a paper-based certification system, which is highly susceptible to food fraud and falsely labelled products. To assure consumers' confidence in organic agriculture, there is a need for an analytical verification process that could differentiate between conventionally and organically produced food.

Working towards this aim, there have been several approaches to clearly identify organically produced plant-based food. All yet known methods, such as transcriptome studies (van Dijk et al. 2012), stabile isotope analysis (Laursen et al. 2014), pesticide residue analysis (Hoefkens et al. 2009), copper chloride crystallisation method (Busscher et al. 2010), or metabolome studies (Bonte et al. 2014), had to face different challenges and could not find their way into practical application. (Capuano et al. 2013) concluded in their review about discrimination and markers to identify organic products that many techniques have been tried but no single one could achieve the given aim yet. A combination of different methods and markers seemed to be the only way to identify falsely labeled products.

Proteome studies in wheat showed an annual stabile difference in protein pattern of flour between the different agricultural systems in two years (Zörb et al. 2009a). Unfortunately, it has been missed to verify the results with multivariate statistics.

In general, a lower protein content of organic cereals could be measured on average over several years (Magkos et al. 2003). This refers to the fact that the mineralisation rate of organically derived nitrogen depends mainly on highly fluctuating weather conditions. But in some years, the protein content has been found as equal. Another study, also realized by Zörb et al. (2009b) showed differences in the metabolic fingerprint including proteins in ears but not in mature grains, again without multivariate statistics. For potatoes, the gap between the protein content of diverse agricultural practices could also be led back mainly to fertilizer management (Lehesranta et al. 2007).

The analysis of the transcriptome focussing on food quality assessment revealed contradicting results: The analysis of different chemical pathways in potato tubers combined with different statistical tools of several agricultural combinations revealed differences in starch synthesis, generation of lipoxygenases and biotic stress responses (van Dijk et al. 2012). But in this case, organic fertilization led to higher, whereas organic crop protection was followed by lower expression of lipoxygenase pathway. Just the starch synthase pathway succeeded in the same direction with an increase in organic agriculture traits. Also for wheat, indication of the "organic status" could be monitored with microarray studies where twelve genes responded to the type of fertilizer. In the gene category associated with storage during seed development, the gliadin transcripts in organic grain showed higher abundance (Lu et al. 2005). In contrast to this, Cicatelli et al. (2014) could not find any variations in the transcriptome using compost and mineral fertilization in potatoes. High-throughput gas chromatography/mass spectrometry technique (GC/MST) with eleven wheat cultivars revealed statistically significant difference of metabolites such as myo-Inositol, 4-aminobutanoate and tryptophan (Bonte et al. 2014), which is a promising result on the way to unambiguously identify organic food. Especially the diversity of potential markers within the metabolome enhanced the potential to find reliable patterns. It has been recently concluded that these findings could be regarded as systematic impact of agricultural systems on the metabolome of plants (Mie et al. 2014).

The reason for differences in proteome, transcriptome or metabolic composition relies on the expression status of genes. An even richer source of biomarkers could be the epigenetic “packaging” as basis of different gene expression. Lopez and Wilkinson (2015) highlighted the potential of epigenetics for crop production and food quality. They concluded that epigenetic information will have a large impact on the future agriculture. Epi-breeding could adapt plants to stressful environments like salt or drought stress and analysis of epigenome during growth could monitor pathogen stress or nutrient deficit earlier than actual methods. The proposed analysis of quality traits of food, using epigenetic fingerprinting has not been tried so far even though many hints like transcriptome and metabolome studies are around.

This knowledge gap was the motivation for our study. We chose cytosine DNA methylation as epigenetic trait for our analysis as it is one of the best-studied epigenetic marks (Koziol et al. 2016). Furthermore, the methods for genome-wide detection of DNA methylation is relatively well established and the here described findings can be of large importance for potato breeding.

Evolved as a recognition system for foreign DNA, adding methyl groups onto DNA bases today usually represses gene expression, and, therefore acts as one of several DNA expression control systems, dependent on methylation context, position and function of the surrounding region.

DNA methylation is regulated by three main pathways: *de novo* methylation (I), maintenance of methylation (II) and active removal (III) of methylation marks (Matzke et al. 2007). In contrast to other eukaryotes, DNA cytosine methylation in plants appears in three sequence contexts: symmetric CpG, CpHpG and asymmetric CpHpH, where H stands for A, C, or T. The *de novo* DNA methylation is associated with the accumulation of 24nt-siRNAs homologous to the genomic region going to be methylated. The mechanism is named as RNA-directed DNA methylation (RdDM, (Wassenegger et al. 1994; Chan et al. 2005; Matzke and Birchler 2005). The RdDM mechanism bears methylated cytosines irrespective of their sequence context. This mechanism is plant-specific and involved in various regulatory processes. DNA methylation acquired by RdDM can be associated with transcriptional gene silencing (TGS) (Castel and Martienssen 2013), genome stability by transposon inactivation (Bucher et al. 2012), hybrid vigour (Shen et al. 2012) and DNA virus defense (Wang et al. 2012).

The methylation status of DNA can be highly variable, but it is considered stabile enough to be persistent through the development of a plant and can be inherited into the next generation (Niederhuth and Schmitz 2014). Environmental stress was found to lead to alternated methylation patterns (Verhoeven et al. 2010; Boyko and Kovalchuk 2011).

We assume differences of methylation in plant defence, hormone and metabolic pathway-related genes. Plant defence and hormone-signalling plays an important role in plant-plant, plant-pathogen and plant symbiosis interactions. Because of the absence of synthetic pesticides in organic agriculture, the plants have to protect themselves, and have to communicate with other plants and beneficial organisms about pathogen attack.

Looking at epigenetic patterns in the context of methylome differences in distinct growth systems could help to promote the organic agriculture, prevent food fraud, help breeders to develop environmentally adapted plants and could be taken as basis of future tools to monitor plant development and health. We assume that the reasons for differences in terms of yield such as different plant protection and fertilization, between organically and conventionally grown plants, also lead to specific adaptations within the DNA methylation status within the plant genome. These differences are further inherited to progeny generations and could be found as epigenetic fingerprint in seeds and other harvested products.

Objectives of this study are: i) first examination of DNA methylation patterns in bio-organic potato (*Solanum tuberosum*) samples in contrast to conventionally grown plants. ii) evaluation of the capability to include DNA methylation analysis in the field of food/product authentication. iii) comparison of differentially methylated regions to get a first impression about the temporal variability of differential methylated regions (DMR).

5.3. Material and Methods

5.3.1. Material

To know the origin and the production system of the potato (*Solanum tuberosum var. Desirée*) samples, we requested them from the DOK-field trial (D: bio-dynamic, O: bio-organic, K: conventional) in Therwil, Switzerland of Agroscope, the Swiss Centre of Excellence for Agricultural Research and the Research Institute of Organic Agriculture (Foschungsinstitut für biologischen Landbau - FIBL). The DOK is one of the largest and the longest lasting long-term field trial of its kind, initiated in 1978. The samples have been kindly provided by Agroscope.

Fully developed tubers (according BBCH Scale 909) were collected at harvest time (Hack et al. 1993) from potato plants (*Solanum tuberosum cv. Desiree*) in four replicates of the bio-organic cropping system (*organic*) and the conventional cropping system (*conventional*) from four years (2001, 2009, 2011, 2012). The complete tubers formed a representative sample of the plot per biological replicate. They were washed, cut, dried, ground, and stored at 8 °C. The bio-organic system is managed according to the guidelines of BioSuisse (Bio-Suisse 2012). The conventional mixed system with mineral fertilisers plus farmyard manure is managed in accordance with the Swiss integrated management standard since 1985 (IP-Suisse). Genetical integrity has been kept as every year new "seed" tubers have been used from the same batch for organic and conventional plots.

We have chosen the sample material and sample years based on:

i) cultivar homogeneity (variety/accession) of minimum 4 years
ii) common and highest available fertilization quantity (1,4 livestock manure units/ha)
iii) most recent sample years

The potato plants were grown in-field that underwent seven-field crop rotation which comprises five field crops: clover grass ley, maize, soybean, winter wheat and potatoes. The crop rotation changed slightly at the end of each rotation period but was identical for all systems. The experimental design is a split-split-plot Latin rectangle with four field replicates. In addition, each crop rotation is replicated three times and arranged in a shifted design with different crops growing in parallel in respective years (A, B, C – table 5). Details can be found in Mayer et al. (2015). Hence, the preceding crops to grown potatoes were different. In 2001 it was clover-grass ley and in 2009, 2011 and 2012 it has been soybean (*Glycine max*).

While under organic farming cultivation conditions, copper-containing fungicides were applied, none of them were applied under conventional growing conditions. The in-field copper and fungicide application has been split into 5-7 applications. The applied substances and amounts of applied copper on the field are listed in table 5.

Table 5: Amount and type of copper application in the DOK- long-term field trial bio-organic fields as well as the fungicide applied in the conventional fields respectively. The DOK field trial has been set-up in three shifted crop rotation replicates [A, B, C] to guarantee a relatively regular sample and species availability. Each year "contains" 4 true biological repetitions on separate plots in the DOK field trial.

Year	Crop rotation replicate	Total amount of copper	Type of copper application/medium (bio-organic fields)	Fungicides applied in the conventionally managed fields
2001	C	4.0 kg Cu ha^{-1}	Copper-Oxychloride (Kupfer 50)	Mapro, Rover Combi, Ridomil
2009	A	4.0 kg Cu ha^{-1}	Copperhydroxide (Kocide DF)	Ridomil Gold, Mapro, Daconil Combi, Revus M2, Valbon, Ranman A + B
2011	C	3.9 kg Cu ha^{-1}	Copperhydroxide (Kocide opti)	Revus, Ridomil, Valbon, Daconil combi DF, Acrobat
2012	B	3.9 kg Cu ha^{-1}	Copperhydroxide (Kocide opti)	Revus MZ, Ridomil Gold, Valbon, Ranman A + B

Several other approaches have used the DOK-field trial in order to identify organic food looking on wheat (*Triticum aestivum*) using a biocrystallization (Kahl et al. 2008), a proteomics (Zörb et al. 2009b) and a metabolomics approach (Bonte et al. 2014).

5.3.2. DNA extraction and Bisulfite Conversion

DNA was extracted from stored samples using a modified column-based DNeasy kit according to manufacturer (Qiagen) protocol, including a first phenol purification step with 100 mg of the grounded potato sample. Integrity of DNA was tested by gel electrophoresis and Ethidiumbromid staining (Supplement 1). Note that fragmented DNA was detectable similar to apoptosis-like laddering indicating degradation/fragmentation due to storage. For analysis of the whole genome methylation approach, the DNA of the samples from 2001 and 2012 in the first three replicates was subjected to a Whole-genome-Bisulfite Sequencing (WGBS) procedure using a post-bisulfite adapter-tagging protocol (PBAT) to account for low amounts of DNA using the EZ DNA Methylation Gold-™ Kit, Zymo Research. DNA concentration after bisulfite conversion was quantified using Qubit and q-PCR.

Whole-genome bisulfite sequencing is able to detect methylated cytosines within the whole DNA pool on a single nucleotide level. During DNA bisulfite reaction, unmethyated cytosines are converted to uracil (Frommer et al. 1992). This occurs via cytosine sulphonate and a following hydrolytic deamination to uracil sulphonate and further a desulphonation step to uracil. Due to the methyl group, 5-methylcytosine is not affected by this transformation. The subsequent polymerase chain reaction (PCR) leads to a C/T-polymorphism which could be quantified and afterwards interpreted as proportion of the original methylation at a specific site by comparing of reads with the original strand or a reference genome.

5.3.3. Library Preparation

Library preparation and sequencing was performed at Novogene Inc. (Novogene Co., Ltd. Novogene Biotech Co., Ltd. Building 301, Zone A10 Jiuxianqiao North Road, Chaoyang District, Beijing). After transfer of samples integrity, quantity and quality of DNA was controlled. Afterwards the DNA has been sheared using Covaris S220 to fragments of 200-400 bp and further processed according the library preparation protocol including terminal repair, A-ligation and ligation of sequencing adapters. HiSeq X Ten (Illumina Inc., 9885 Towne Centre Drive. San Diego, CA 92121 U.S.A.) sequencing, with paired-end 150 bp, including multiplexing of other, non-bisulfite treated, samples has been applied to account for the unbalanced base ratio in WGBS libraries. The effective concentration of the final library has been >2 nM. A Lambda Phi X spike-in with 0.5% of DNA weight has been applied as conversion standard. The demultiplexed raw data has been received.

5.3.4. Bioinformatics

The entire pipeline has been written in snakemake (Köster and Rahmann 2012), using conda environments (Anaconda Software Distribution 2016) and is depicted in Supplement 2. If not stated otherwise, the default tool parameter settings have been used.

Remaining adapter and faulty sequences of the raw reads have been carefully analysed with FastQC (Leggett et al. 2013) and removed during trimming using TrimGalore (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) [--illumina –paired -q 20 --clip_r2 18 --clip_r1 7 --three_prime_clip_R1 18 --three_prime_clip_R2 3]. We applied Bismark (Krueger and Andrews 2011) with bowtie2 (Langmead and Salzberg 2012) [--score_min L,0,-0.6] for mapping of each technical replicate separately including extensive quality check of the .bam files using samtools (Li et al. 2009), and bamqc from the qualimap package (García-Alcalde et al. 2012). Afterwards, the technical replicates have been merged by biological replicates and sorted with samtools prior to deduplication with bismark de-duplication function. Also the calculation of methylation levels for each cytosine has been performed with the bismark package. Calling of differentially methylated cytosines (DMC) and DMR has been performed with Defiant (Condon et al. 2018) in five different parameter setting combinations [defiant settings: a) minimal coverage: 10, CpN 5, $p < 0.05$, 30% methylation difference minimum; b) minimal coverage: 10, CpN 5, $p < 0.05$, 10% methylation difference minimum; c) minimal coverage 10, CpN 1, $p < 0.05$, 10% methylation difference minimum; d) minimal coverage 10, CpN 1, $p < 0.05$, 30% methylation difference minimum; e) minimal coverage 10, CpN 1, $p < 0.05$, 50% methylation difference minimum]. Visualization has been accomplished using ViewBS (Huang et al. 2018), IGV (Robinson et al. 2011) and R.

To evaluate the biological relevance of our findings, Gene-Ontology analysis (GO) were performed, based on differentially methylated cytosine position (DMC) [defiant settings: d)] in relation to known genes for hypermethylated (less in conventional samples) and hypomethylated (less in organic samples) DMCs separately. If a DMC has been spotted within a 2000 nucleotide window upside the transcription start site (TSS), we took the related gene together with its functional annotation and computed the distance to a background of all known coding genes using PLAZA 4.5 (van Bel et al. 2018). As reference genome, the *Solanum tuberosum* annotation 3.0.43, downloaded from the Ensemble Plants Database (http://plants.ensembl.org/Solanum_tuberosum/Info/Index) has been used.

All scripts are available on request.

5.3.5. Copper analysis

The total copper content of all samples has been analyzed using high pressure digestion [150 mg sample material in 1 ml 65% HNO_3, 8 h at 170 °C, adding 9 ml distilled water]. The measurement was done using Inductively-Coupled Plasma-Atomic-Emission-Spectrometry (ICP-AES) with upstream pressure digestion.

5.4. Results

5.4.1. Whole genome bisulfite sequencing of potato samples

Supplement 2 shows the bioinformatics pipeline with each node being a rule and each line showing the handover of datasets. In Table 6, the raw output of reads generated for each sample is given. Analyzed were samples from two years (2001 and 2012) with three biological replicates each from both growing conditions (K conventional, O organic). Table 6 shows the resulting coverage per biological replicate, the duplication level and the conversion rate (S3a: MultiQC-Report with the number of reads per sequencing run/pseudo-technical replicate).

Table 6: Mean coverage, number of reads prior and after deduplication and conversion rate per biological replicate.

	Mean Coverage after mapping (fold)	dedup_leftover (number of reads)	total prior to dedup (number of reads)	Conversion Rate
K_52_2001_1	66,0416	215898398 (83.42 % of total)	258815030	0,9949
K_53_2001_2	62,588	204519192 (84.97 % of total)	240697340	0,9948
K_54_2001_3	37,2491	120466740 (58.44 % of total)	206146508	0,9956
K_70_2012_1	45,6284	147623939 (71.35 % of total)	206905121	0,9947
K_71_2012_2	54,5479	176419142 (66.25 % of total)	266298578	0,9949
K_72_2012_3	37,909	122602672 (68.44 % of total)	179148320	0,9959
O_49_2001_1	44,6694	144446420 (63.02 % of total)	229224302	0,9954
O_50_2001_2	80,654	265836596 (78.97 % of total)	336649010	0,9962
O_51_2001_3	58,1251	190057823 (79.77 % of total)	238264224	0,9947
O_67_2012_1	52,764	170638483 (70.45 % of total)	242199308	0,9958
O_68_2012_2	38,4213	125369199 (80.25 % of total)	156215284	0,9961
O_69_2012_3	49,164	158979415 (69.47 % of total)	228850810	0,9949

The coverage distribution over the position within the potato genome (*Solanum tuberosum*, Soltub 3.043) could be examined in Figure 14.

The global methylation level per C context (CG, CHG, CHH) and biological replicate is depicted in Figure 15 (S3b MultiQC-Report with M-bias plot, S4-S6 methylation level distribution per sample and C-context, S7 Exemplary Methylation Level Distribution over the genome). The detected level of methylation in the three sequence contexts CG ~75%, CHG ~50% and CHH 25% agrees with previous reports for potato (Wang et al. 2018). This result meets the expectations as symmetric methylation is accumulating by the maintenance mechanism and CHH methylation via the de novo methylation pathway of RdDM. Two groups of samples are detectable, showing similar differences in overall methylation. Looking at the distribution of the methylation level over the genome, it can be seen that especially highly methylated regions are affected at the end of chromosomes, which might be indicative for pronounced degradation in some of the samples (Supplement 1).

To examine the methylation level in more detail, we calculated the methylation level in relation to known genes, repeat regions, coding sequence (CDS), exons, 3´-untranslated regions (UTR), 5´UTRs and inter-genic regions using a length normalization to account for different length of genomic features including 2 kb prior and after the feature (Figure 16). The positions of features have been taken from the *Solanum tuberosum* annotation 3.0.43. For repeats, we used the positions described in Mehra et al. (2015).

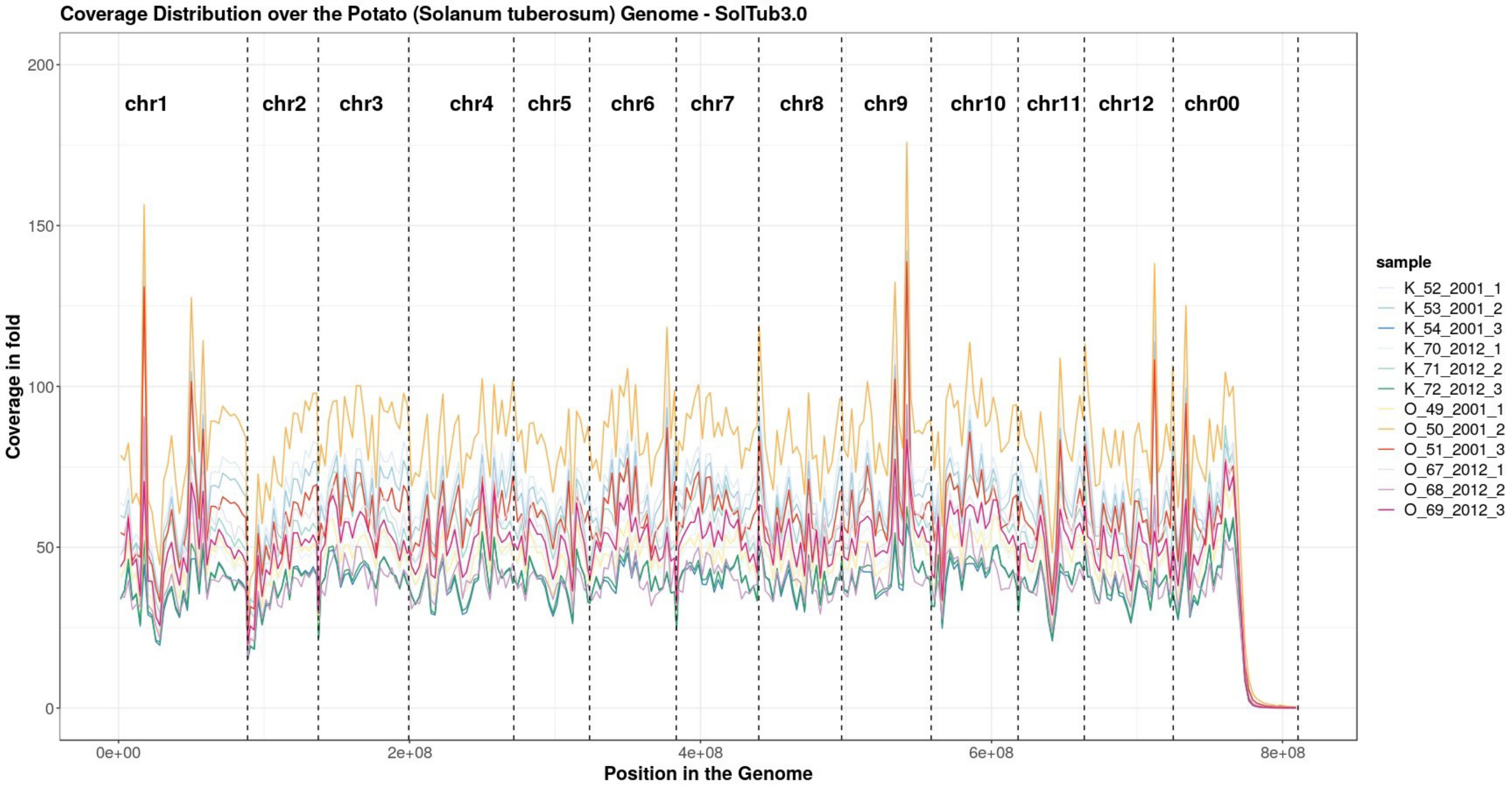


Figure 14: Coverage Distribution over the potato (Solanum tuberosum) reference genome Sol.Tub3.0 for all samples analysed by WGBS, dotted lines show the border of chromosomes. Coverage is depicted in "fold" in relation to the position in the genome.

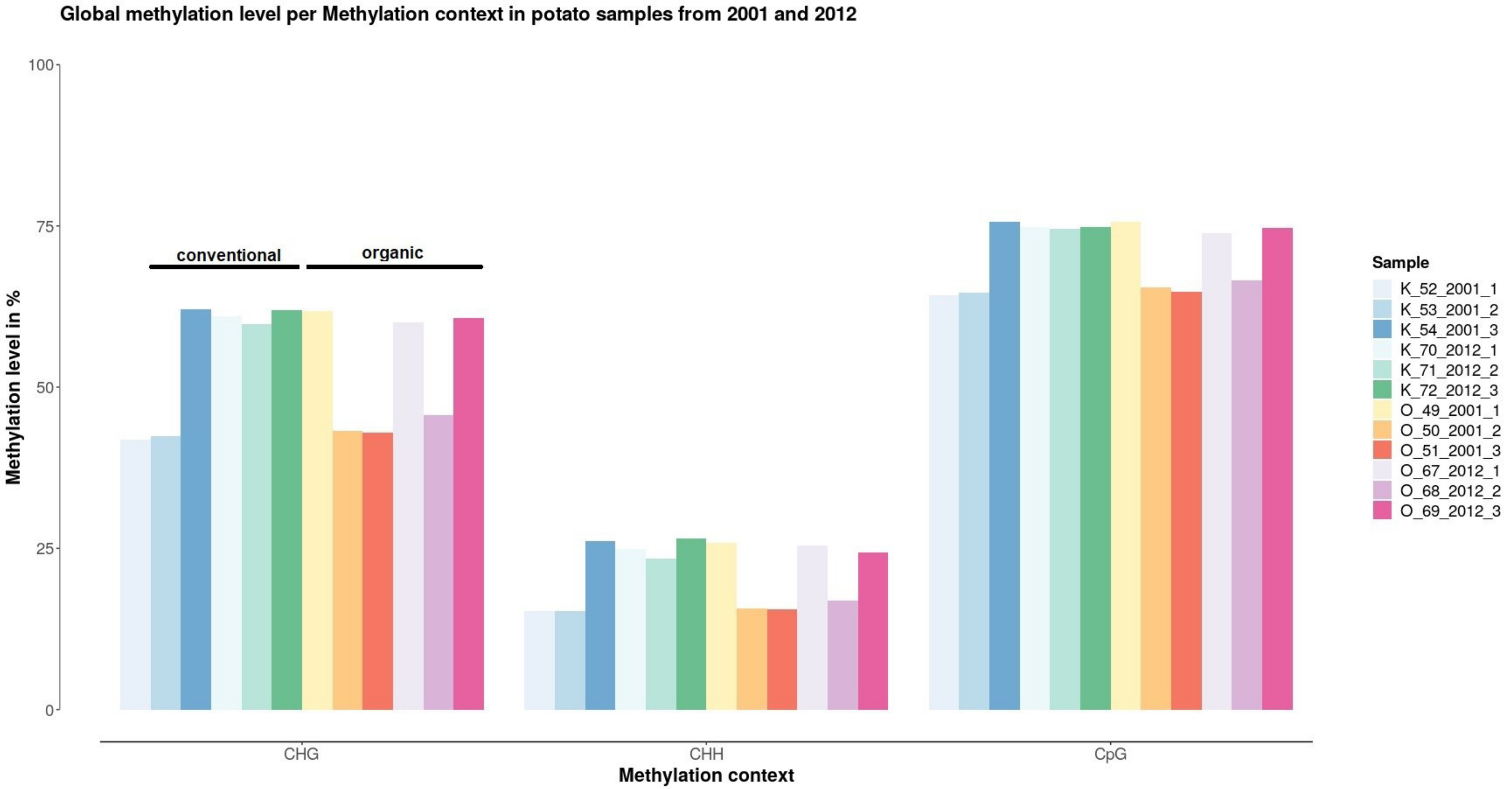


Figure 15: Global methylation level per sample and methylation context. Cytosine methylation could occur in different contexts: symmetric CpG, CpHpG and asymmetric CpHpH, where H stands for A, C, or T.

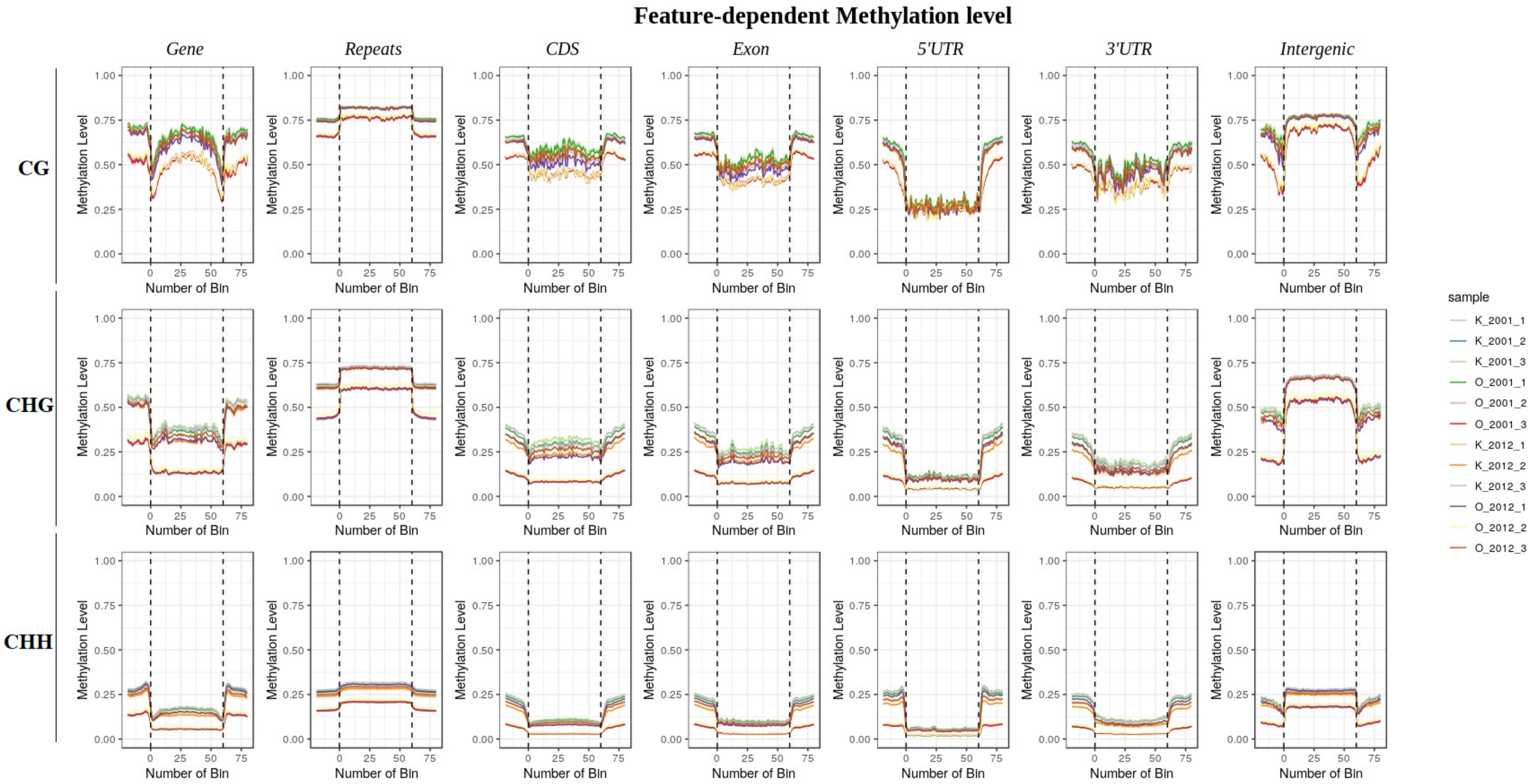


Figure 16: Feature-dependent methylation levels prior, within and outside of genomic features splitted into different methylation context (CG, CHG and CHH). A binning normalization has been used to account for different feature length within the potato (Solanum tuberosum) genome with 2 kb prior and 2 kb after the respecti ve feature.

5.4.2.Identification of differentially methylated regions

DMR were identified using Defiant in five different parameter setting combinations. The number of identified DMR/DMC are shown in Table 7, including the number of overlapping DMR/DMC between the two years examined by PBAT-WGBS. The DMR calling output is displayed exemplary for a) in form of tables for each examined year (S8 & S9: List of DMR for 2001 and 2012). Two of the identified DMR have high biological and agronomical relevance and are shown in figure 17 (SUT1) and figure 18 (ATOX1).

Table 7: Number of differentially methylated regions (DMR) or cytosines (DMC) examined between organic and conventionally grown potato samples in five different parameter setting combinations including the number of overlapping DMR/DMC between the two years.

	a)	**b)**	**c)**	**d)**	**e)**
	Coverage: 10 CpN: 5 p < 0.05 MethDiff: 30%	Coverage: 5 CpN: 5 p < 0.05 MethDiff: 10%	Coverage: 5 CpN: 1 p < 0.05 MethDiff: 10%	Coverage: 10 CpN: 1 p < 0.05 MethDiff: 30%	Coverage: 10 CpN: 1 p < 0.05 MethDiff: 50%
2001	40	7784	2329602	13922	251
overlap	0	60	91565	6	0
2012	9	19291	2924040	13924	17

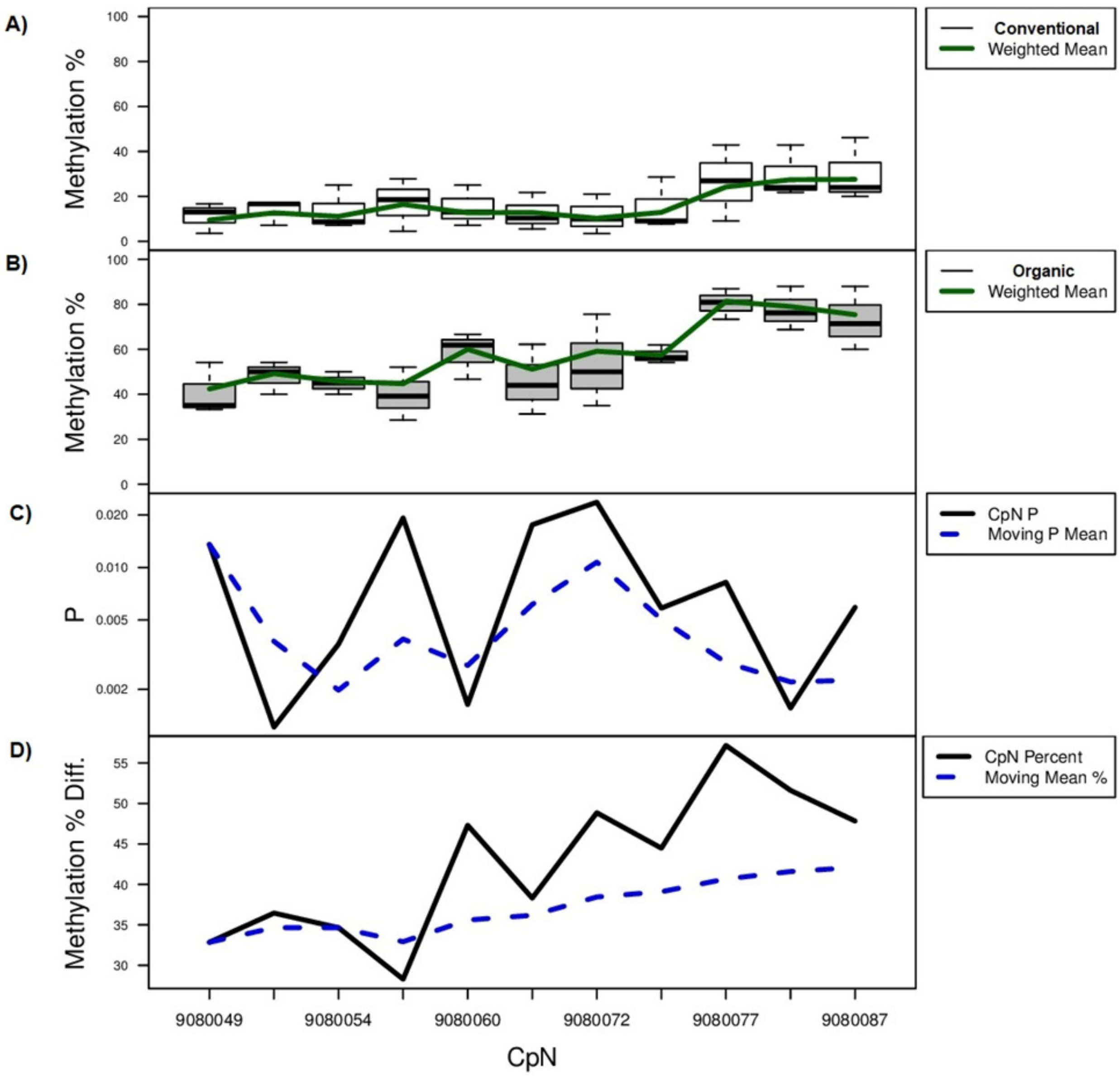


Figure 17: Differentially methylated region associated with SUT1 (depicted and called by Defiant). A) represents the conventional potato samples. As "Case" we used the organic samples shown in B). The two upper windows (A&B) show the methylation level per cytosine position in the genome with standard derivation within the group of organic or conventional samples. The third window (C) shows p-value per position, and D) shows the methylation difference between A) and B). This DMR has been called for samples from 2001.

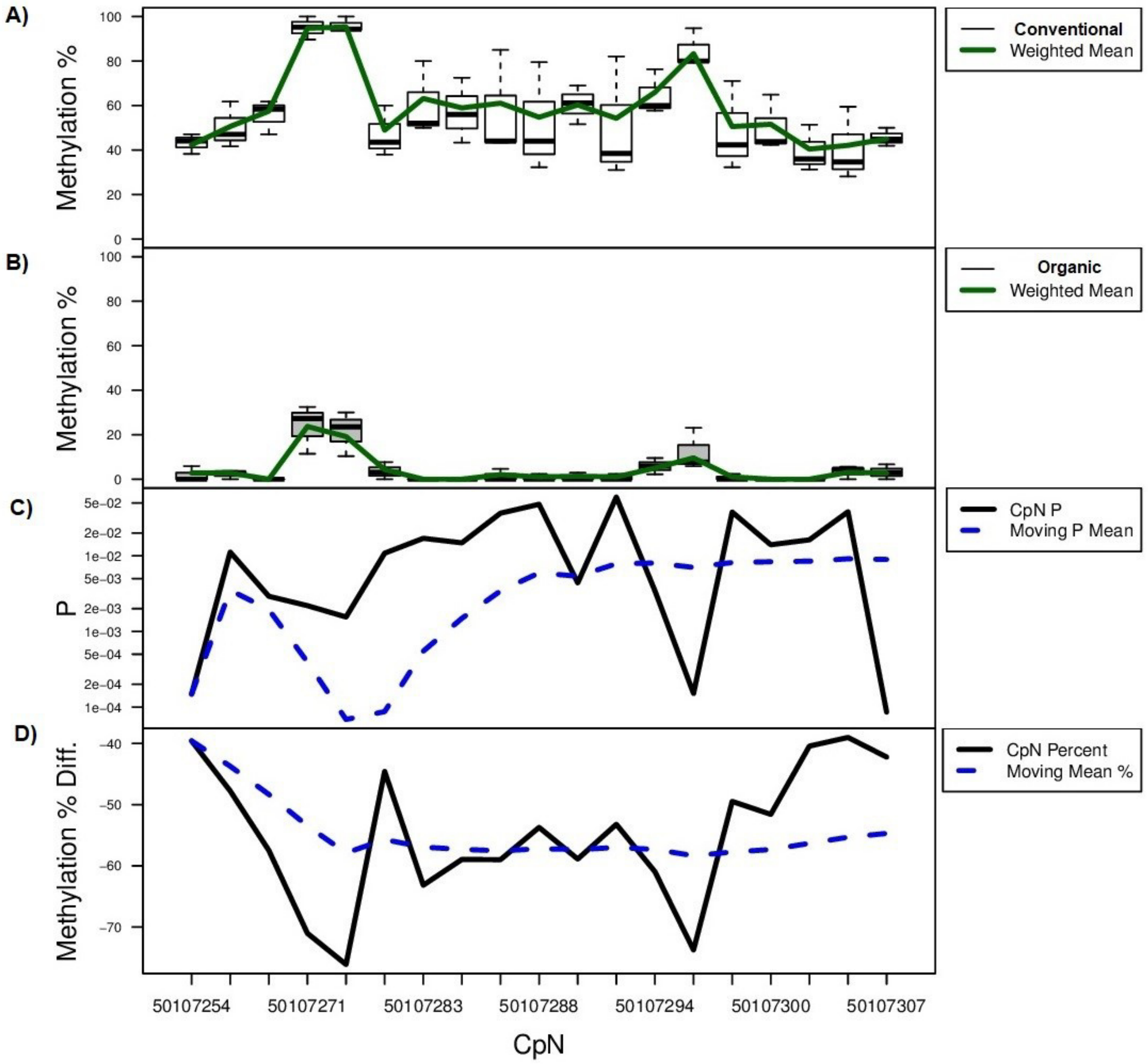


Figure 18: Differentially methylated region associated with ATOX1 (depicted and called by Defiant). A) represents the conventional potato samples. As "Case" we used the organic samples shown in B). The two upper windows (A&B) show the methylation level per cytosine position in the genome with standard derivation within the group of organic or conventional samples. The third window (C) shows p-value per position, and D) shows the methylation difference between A) and B). This DMR has been called for samples from 2001.

Among the 60 overlapping DMR (Table 7 b), 6 are found to be located in coding regions of annotated genes (S10). These are of interest as DNA methylation within coding genes can be associated with post transcriptional gene silencing. Three of these showed the change in methylation level in the same direction: PGSC0003DMT400034908 annotated as MDR-like ABC transporter (2001: -18.4%, 2012: -16.6%), PGSC0003DMT400077180 (2001: -12.7%, 2012: -22.7%) annotated as Vacuole sorting receptor protein PV72 and PGSC0003DMT400000932 (2001: -15.7%, 2012: -17.3%) annotated as MYB-Transcription factor. A negative overall methylation difference of the DMR indicates a demethylation in the organic samples.

The results of GO term analysis are depicted in figure 19 for hypermethylated DMCs in 2012. Other GO term analysis results can be found in S11 as tables for biological process (BF) and molecular function (MF) separately. For GO term analysis we focussed on DMC calling settings d). The findings suggest a significant enrichment [p<0.05] of DMC in 2012 in gene-promoter regions known to be involved in binding processes compared to the background of all known coding genes. In 2001 and 2012 for hyper- and hypomethylated DMCs in promoter regions, we see a high proportion of genes to be involved in metabolic processes, nitrogen compound metabolic process or response to stimuli pathways. 6 DMCs have been found in both years to be differentially methylated.

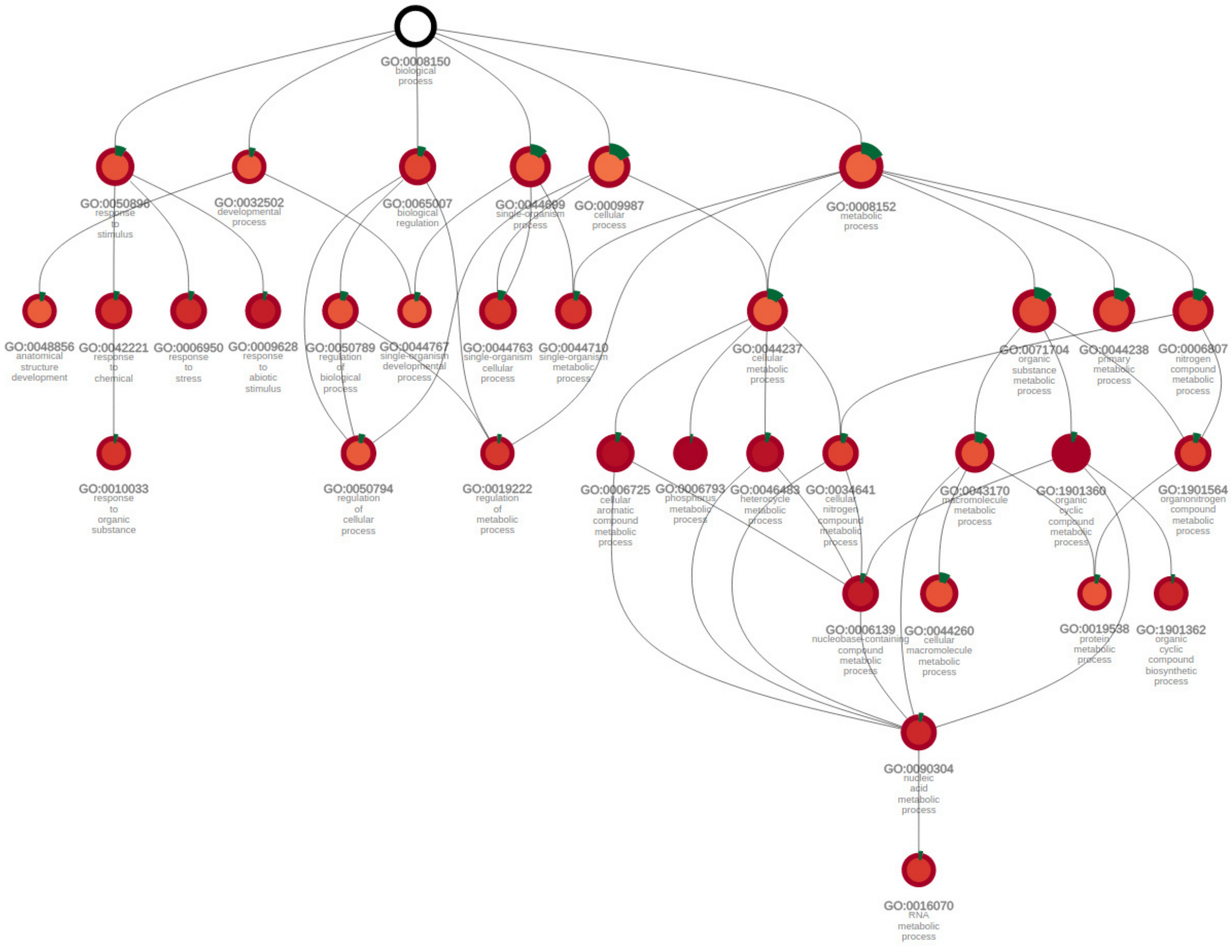

GO:0008150 biological process
GO:0050896 response to stimulus
GO:0032502 developmental process
GO:0065007 biological regulation
GO:0044699 single-organism process
GO:0009987 cellular process
GO:0008152 metabolic process
GO:0048856 anatomical structure development
GO:0042221 response to chemical
GO:0006950 response to stress
GO:0009628 response to abiotic stimulus
GO:0050789 regulation of biological process
GO:0044767 single-organism developmental process
GO:0044763 single-organism cellular process
GO:0044710 single-organism metabolic process
GO:0044237 cellular metabolic process
GO:0071704 organic substance metabolic process
GO:0044238 primary metabolic process
GO:0006807 nitrogen compound metabolic process
GO:0010033 response to organic substance
GO:0050794 regulation of cellular process
GO:0019222 regulation of metabolic process
GO:0006725 cellular aromatic compound metabolic process
GO:0006793 phosphorus metabolic process
GO:0046483 heterocycle metabolic process
GO:0034641 cellular nitrogen compound metabolic process
GO:0043170 macromolecule metabolic process
GO:1901360 organic cyclic compound metabolic process
GO:1901564 organonitrogen compound metabolic process
GO:0006139 nucleobase-containing compound metabolic process
GO:0044260 cellular macromolecule metabolic process
GO:0019538 protein metabolic process
GO:1901362 organic cyclic compound biosynthetic process
GO:0090304 nucleic acid metabolic process
GO:0016070 RNA metabolic process

Figure 19: Gene-Ontology-term analysis based on differentially methylated cytosines (coverage min. 30x, 30% methylation difference) in 2 kb promoter regions prior to genes, exemplarily for hypermethylated DMCs in 2012 grouped by biological processes. Only significant (p<0.05) enrichments in relation to all known coding genes in the potato genome have been used.

5.4.3. Copper content

Total copper content of potato tubers is shown in figure 20. Our analysis reveals no significant difference in the copper content in the dried potato tuber samples grown under conventional and organic farming conditions. Among the different years minor differences are detectable.

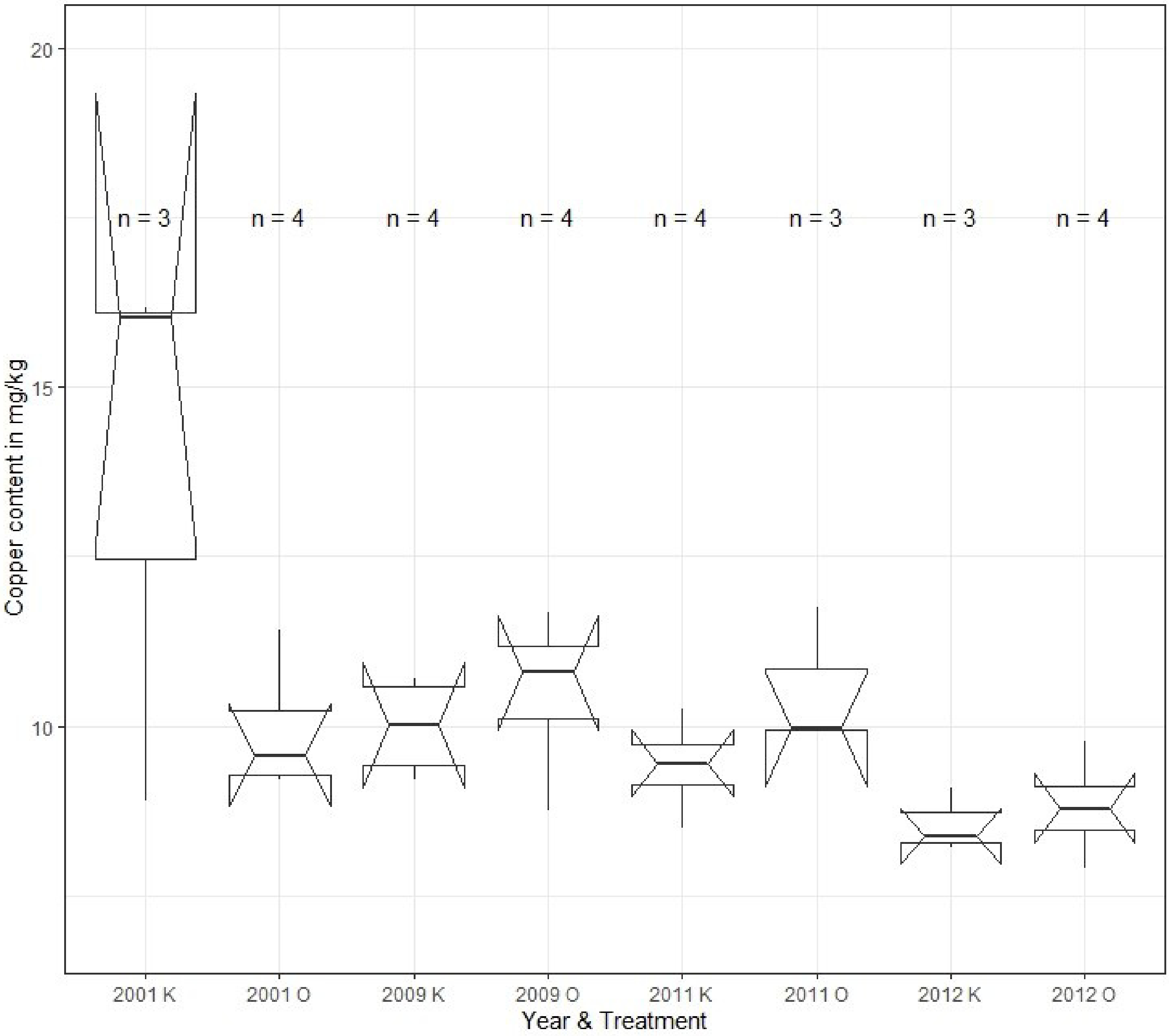


Figure 20: Total copper content in potato tuber measured by ICP-AES after high pressure nitric acid digestion for different years and treatments (organic = O, conventional = K), notched box plots show sample distribution, while we are aware of the low sample size (n=3 or n=4). Different sample sizes occurred due to mislabelling of samples after harvest. All uncertain samples have been removed.

5.5. Discussion

As mentioned earlier, the variability of nutrients and plant metabolites depends mainly on species variety, growing conditions and time. This variability also influences DNA methylation. A limitation of WGBS is that bisulfite cannot differ between 5-hydroxymethylcytosine and 5-methylcytosin. This has to be taken into account but recent observations concluded a negligible relevance of 5-hydroxymethylcytosine in plants (Erdmann et al. 2014).

Compared to Wang et al. (2018), a similar level and distribution of DNA methylation was determined. Although, different cultivars were used (*Solanum tuberosum cv. Pacific* vs. *Desirée*). The overall methylation level in this study has been ~0.7 for CG, ~0.41 for CHG and 0.15 for CHH methylation. The methylation level in our study ranged between 0.64 – 0.76 for CG, 0.42 – 0.62 for CHG, 0.15 – 0.27 for CHH methylation context. In terms of methylation structure, Wang et al. (2018) also reported an increase of methylation in transposable elements as well as a decrease for CHG and CHH gene body methylation but also for transcription start site methylation.

We found a high number of stable methylated regions being mainly associated with repetitive elements and being hardly modified according to environmental changes. A much lower number of methylated regions showed differences in their methylation level. These regions were termed differentially methylated regions (DMR) and they were sorted according the farming type. A standardized definition of a DMR remains a controversial discussion in the field of epigenetics. To highlight the impact of parameter settings on DMR/DMC calling and overlapping DMR/DMC between the two years, we decided to use five different setting combinations. As these DMR underwent changes based on the environment, we consider them as meta-stable (Fujimoto et al. 2012). Based on our analysis, the number of regions with a consistent methylation pattern under organic and conventional farming was relatively low due to our approach to minimize the false positive rate.

In total, 49 DMR (40 in 2001 and 9 in 2012) were identified between the farming systems (DMR calling set a). 21 of 49 where hypomethylated and 28 hypermethylated. Thirteen DMR are associated with potential promoter regions (1000 bp cut-off) and might have an effect on the transcriptional rate of their downstream genes (Havecker et al. 2012). Five of them are associated with coding regions and the methylation here might be a consequence of silencing events. This has been the most stringent DMR calling setting with no overlap between the two years. The other DMR are associated with unspecific intergenic regions and non-coding genomic features. As this list of detected DMR is very short and the detected variation in DMR is relatively high, it should be considered that the identified DMR are found by random, whereas the strong differences in all replicates analysed and low p-values indicate their reliability.

SUT1 (PGSC0003DMG400009213) encodes a sucrose transport gene and is located at chromosome 11. The identified DMR (chromosome 11; 9080049-9080127) has a length of ~78 nucleotides and contains ~20 cytosines (dependent on the settings of the DMR calling). WGBS results showed high methylation levels in association with organic growth conditions for the year 2001.

At chromosome 7, a DMR (50107254-50107307) was identified, located in the promoter of a gene annotated as ATOX1 (PGSC0003DMG400020423). ATOX1 encodes a copper-binding metal-chaperone-protein involved in copper transport (Blaby-Haas et al. 2014). Its molecular function and association to diseases in mammals was strongly investigated (Ge et al. 2019). This DMR (Figure 18) was identified in the year 2001 samples to be specifically methylated in conventionally grown tubers and significantly less/hypo-methylated in organic tubers. As this DMR was located in the promoter region of the gene, it might be associated with transcriptional gene regulation of the corresponding gene (Havecker et al. 2012). This regulation should be addressed in future experiments. The missing association of ATOX1 with a similar DMC/DMR in 2012 could have multiple reasons. For example, in 2001 copper-oxychloride as fungicide has been used whereas 2012 copper-hydroxide has been applied.

As organic farming is restricted to copper-containing fungicides, which were also applied under the experimental organic conditions (Table 5) but not under the conventional treatment, we speculate that the transcriptional regulation of ATOX1 is activated by the presence of copper. This regulation might be associated with demethylation under presence of copper or methylation in the absence of copper. However, as the material has been dried, grounded and stored for several years transcriptional analysis will be difficult and could not be covered in this publication.

This is no contradiction to the fact that the copper content in all investigated potato tubers was not significantly different because copper fungicides are applied to the soil and or plant surface, while copper uptake into tubers occurs mainly from soil not exposed to copper fungicide. In addition, the fungicial action of copper is effective despite its general low application amount of a few kilograms per hectare not significantly influencing copper uptake into plants.

Looking at other DMR/DMC calling settings we see 60 overlapping DMR for a minimal coverage of 5-fold and 10% methylation difference for each cytosine in the respective DMR (DMR calling set b). Also looking at DMCs we find overlapping positions for example 91.565 DMC for DMR calling set c with a coverage of 5-fold per cytosine minimum, CpN 1, $p < 0.05$, and 10% methylation difference between the organic and conventional treatment.

Three genes out of the set of 60 with DMR inside the coding regions were identified in both years to be differentially methylated in the same direction. Here, the Multi-drug ABC transporter is of particular interest as this transporter family has been reported to be not only involved in the detoxification of xenobiotics (Remy and Duque 2014), but also in lateral root development and root hair formation (Santelia et al. 2005). Furthermore, the gene encoding the PV72 Vacuolar Sorting Receptor was also found to be associated with a DMR. PV72 has been described to be involved in the transport of storage protein precursors (Shimada et al. 2002) but also has characteristics of a vacuole sorting receptor (Watanabe et al. 2004). The third gene (PGSC0003DMT400033123) encodes Cytokinin oxidase/dehydrogenase 1. Expression of CKX in barley during germination was reported to be influenced by CKX2.1 promotor methylation and correlated to drought stress during grain establishment (Surdonja et al. 2017).

Interestingly, several of the identified DMR are directly associated or adjacent to genes involved in transport processes (e. g. PGSC0003DMT400064008, PGSC0003DMT400021790, PGSC0003DMT400089747, PGSC0003DMT400077180).

5.6. Conclusions

The advantage of this novel epigenetic authentication method is that we do not focus on just one agronomical trait, which characterizes the organic cultivation conditions, e. g. fertilization or pest control. Using this method, we could reveal differences within the epigenetic fingerprint in terms of fertilization beside plant protection but also in terms of different bacteria/fungi ratio and differences in biodiversity of soils. Further, this method could be applied towards every plant-based trading good and it is not limited to growing tissue or protein-rich parts of the plant. Furthermore, the number of possible epigenetic markers is just limited by genome size and the occurrence of cytosine in the genome. This enables us to search into an enormous pool for a few different methylated regions.

Beside this, we might answer much more questions using further data obtained from additional analysis, especially about environment-genome interactions. It is important for future breeding strategies to get information about the "plants view" on their environment and the heredity of agriculturally relevant, obtained features.

It is still a matter of debate, whether the variation between the cultivars and the years are greater or lower than the variation between the agricultural systems (Laidig et al. 2008). But recently there are some hints that for only a few metabolites the differences within the systems predominate (Bonte et al. 2014).

This topic is especially important because of the enormous development in the organic food sector and the need to assure high quality and authenticity of sold products to maintain consumers' confidence. To confirm our results, it is necessary to verify them in further growing years, growing sites and with several varieties.

For future research, a transfer of the developed method to other agronomically relevant species is possible, even if there is no reference genome available (van Gurp et al. 2016).

5.7. Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation, to any qualified researcher. The main raw datasets are available via the e!Dal Repository of Leibniz Institute of Plant Genetics and Crop Plant Research (IKP) Gatersleben via the following link:

https://doi.ipk-gatersleben.de/DOI/c3cc1902-1f5c-4417-a741-7b44b6bfe366/6a35fa2f-6609-4587-8ccd-0eb1da9e63e7/2/1847940088

5.8. Author contributions

Conceptualization: CG, MK, LA, IG and BG; Methodology: CG and MK; Investigation: CG, MK; Writing—original draft preparation: CG and MK; Writing—review and editing: MK, LA, PM, JM, IG, BG; Visualization: CG; Supervision: IG and BG; Project administration: CG, IG and BG; Funding acquisition: CG, IG and BG

5.9. Funding

This research was funded by the state of Saxony-Anhalt (Graduate Funding for CG) and Volkswagen Stiftung. We acknowledge the financial support within the funding programme Open Access Publishing by the German Research Foundation (DFG).

5.10. Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

5.11. Acknowledgments

We thank all the members of the Bioinformatics Group of Halle-Wittenberg University, especially Samar Fatma, Alexander Gabel, Jan Grau, Silvio Weging, Claus Weinholdt, Ioana Lemnian and Martin Porsch for valuable discussions as well as for support in terms of biostatistics and visualization.

In terms of WGBS experimental support and planning we thank Claude Becker and Etienne Bucher, for copper analysis we thank Klaus Kaiser and Heike Maennicke. Furthermore we thank Sandra Driesslein for DNA extraction.

5.12. Supplementary Material

The Supplementary Material for this article will be made available online directly after publication. Furthermore the Supplementary Material is available under: https://github.com/grehl/WGBSpot

6. Conclusion and summary discussion

Within this thesis, the DNA methylation patterns of potato (*Solanum tuberosum*) samples out of a comparative field trial with organic and conventional plots have been examined to decode the epigenetic fingerprint in a pseudo-natural environment. To approach the topic, to establish DNA methylation analysis at Halle-Wittenberg University and to get familiar with this kind of data it has been started with an intensive literature review (chapter 3), including the planning, set up of experimental design and state of the art of DNA methylation analyses. As the question remained which alignment software should be used best, this topic has been addressed in chapter 4. To finally look at DNA methylation patterns, a large-scale data analysis pipeline has been built to be able to finally analyse the datasets of samples from DOK field trial (biodynamic (D), organic (O) and conventional (K) for German: "konventionell"), a long-term comparison of organic and conventional agriculture systems, which has been initiated in 1978 by Research Institute of Organic Agriculture (FIBL) in Frick, Switzerland. This has been shown in chapter 5.

All in all we find significant DMC and DMR which are indicative for the land use system. However, we still do not reliably know how variable these patterns are. The variability between the different years, varieties or locations has to be addressed in further evaluations.

Concerning the initial hypotheses:

> H1 The agricultural practice under which a plant has been grown could be monitored using analysis of DNA methylation status in harvested products like potato tubers.

This point has been shown with this dissertation and proven to be correct.

> H2 Based on transcriptome and metabolome studies, we expect the methylation differences in plant defence, hormone and metabolic pathway related genome regions as well as in transposal regions.

Transcriptome studies suggested the gene ferritin 1 as well as genes from lipoxygenase, jasmonate pathway (van Dijk et al. 2012), nitrogen metabolism and storage protein synthesis (Lu et al. 2005) as highly promising, which was partly confirmed by Gene Ontology (GO) analysis. We find methylation differences in several genomic pathways. As described in chapter 5 and evaluated based on GO analysis of DMC positions in relation to genes, these differences mainly occur in metabolic pathways (e. g. nitrogen compound metabolic processes or cellular macromolecule metabolic processes) but also in genes related to transport processes within the plant. Defense response is at position 96 of 1337 for 2001 (152 for 2012) hypermethylated DMC, whereas plant hormone related pathways are at position 54 of 1337 (60 for 2012).

The differences in metabolic pathways could be explained with differences in nutrition status and the occurrence of nitrogen mineralisation in organically fertilized soils. The nitrogen uptake takes place with the same enzymes, but the nitrogen and nutrient supply is more constant in conventional agriculture because the mineralisation depends on soil temperature and soil humidity.

> H3 The epigenetic variation of some DNA methylation islands between the years is lower than between the agricultural systems. Therefore, DNA methylation fingerprinting could be used as biomarker to confirm the food production system.

Some DNA methylation islands have shown to be highly variable between the agricultural systems and the years examined. We nevertheless find DMC and DMR to be significantly differentially methylated in the same direction in both years. The usage as biomarker to confirm the food label of specific food production manners nevertheless is questionable at this moment as this field of study is at its beginning. Further, an other approach has to be used for the DNA methylation analysis for practical application in food authentication as WGBS analyses are cost-intensive and contain lots of redundant information. If we know which cytosines are of interest, confirmation could be done by methylation-sensitive restriction enzyme digestion PCR (MSRE-PCR) if the respective restriction enzymes have been chosen. Alternatively rrBS approaches might be of interest also. Especially the combination of DMC calling and MSRE-PCR with machine learning approaches is promising to further approach the aim to distinguish between organic and conventional products using DNA methylation patterns.

All in all the scientific benefit of this dissertation and the related project might be:

- flag-ship project in the field of plant-based product authentication
- going from greenhouse experiments to an agricultural environment using a non-model organism
- detailed methylome information for a rarely epigenetically analysed organism (*Solanum tuberosum*)
- one of the first methylome studies from field grown plants
- identification of novel environmental parameters influencing methylation patterns (e.g. plant protection)
- unravelling novel gene regulatory parthways including epigenetic mechanisms
- high-end approach using PBAT-WGBS including bioinformatics in the range of terabytes
- interdisciplinary approach including epigenetics, genetics, plant physiologies, agricultures and biochemistries perspective
- building, testing and application of a large scale WGBS DNA methylation analysis pipeline
- independent benchmarking study of alignments tools for WGBS datasets out of the perspective of users
- collection of information concerning the state of the art in the field of DNA methylation analysis using WGBS, informative for future PhD candidates, project initiators and people working in the field of epigenetics
- promotion of the research about organic agriculture, prevent food fraud, help breeders to develop environmentally adapted plants and could be taken as basis of future tools to monitor plant development and health

In general, this approach offers the possibilities to monitor the storage, transport, cultivation and processing conditions of all products which contain methylated DNA. Therefore this could be a good tool to monitor and enlighten food fraud, to strengthen consumers´ confidence and to enhance the reliability of peer-to-peer trading in the value chain.

7. Publication bibliography

Aberg, Karolina A.; Chan, Robin F.; Shabalin, Andrey A.; Zhao, Min; Turecki, Gustavo; Staunstrup, Nicklas Heine et al. (2017): A MBD-seq protocol for large-scale methylome-wide studies with (very) low amounts of DNA. In *Epigenetics* 12 (9), pp. 743–750. DOI: 10.1080/15592294.2017.1335849.

Alaru, Maarika; Talgre, Liina; Eremeev, Viacheslav; Tein, Berit; Luik, Anne; Nemvalts, Anu; Loit, Evelin (2014): Crop yields and supply of nitrogen compared in conventional and organic farming systems. In *Agricultural and Food Science* 23, pp. pp. 317–326.

Anaconda Software Distribution (2016): Computer software. Vers. 2-2.4.0. Anaconda, Nov. 2016. Web. <https://anaconda.com>.

Asami, Danny K.; Hong, Yun-Jeong; Barrett, Diane M.; Mitchell, Alyson E. (2003): Comparison of the total phenolic and ascorbic acid content of freeze-dried and air-dried marionberry, strawberry, and corn grown using conventional, organic, and sustainable agricultural practices. In *Journal of agricultural and food chemistry* 51 (5), pp. 1237–1241. DOI: 10.1021/jf020635c.

Ausin, Israel; Feng, Suhua; Yu, Chaowei; Liu, Wanlu; Kuo, Hsuan Yu; Jacobsen, Elise L. et al. (2016): DNA methylome of the 20-gigabase Norway spruce genome. In *Proceedings of the National Academy of Sciences of the United States of America* 113 (50), pp. E8106-E8113. DOI: 10.1073/pnas.1618019113.

Bateman, Alison S.; Kelly, Simon D. (2007): Fertilizer nitrogen isotope signatures. In *Isotopes in environmental and health studies* 43 (3), pp. 237–247. DOI: 10.1080/10256010701550732.

Bateman, Alison S.; Kelly, Simon D.; Jickells, Timothy D. (2005): Nitrogen Isotope Relationships between Crops and Fertilizer. Implications for Using Nitrogen Isotope Analysis as an Indicator of Agricultural Regime. In *J. Agric. Food Chem.* 53 (14), pp. 5760–5765. DOI: 10.1021/jf050374h.

Bio-Suisse (2012): Bio suisse standards for the production, processing and marketing of bud produce from organic farming.

Blaby-Haas, Crysten E.; Padilla-Benavides, Teresita; Stübe, Roland; Argüello, José M.; Merchant, Sabeeha S. (2014): Evolution of a plant-specific copper chaperone family for chloroplast copper homeostasis. In *Proceedings of the National Academy of Sciences of the United States of America* 111 (50), p. E5480-7. DOI: 10.1073/pnas.1421545111.

Bock, Christoph; Tomazou, Eleni M.; Brinkman, Arie B.; Müller, Fabian; Simmer, Femke; Gu, Hongcang et al. (2010): Quantitative comparison of genome-wide DNA methylation mapping technologies. In *Nature biotechnology* 28 (10), pp. 1106–1114. DOI: 10.1038/nbt.1681.

Bonte, Anja; Neuweger, Heiko; Goesmann, Alexander; Thonar, Cecile; Mader, Paul; Langenkamper, Georg; Niehaus, Karsten (2014): Metabolite profiling on wheat grain to enable a distinction of samples from organic and conventional farming systems. In *Journal of the science of food and agriculture* 94 (13), pp. 2605–2612. DOI: 10.1002/jsfa.6566.

Bortoleto, G. G.; De Nadai Fernandes, E. A.; Tagliaferro, F. S.; Ferrari, A. A.; Bueno, M. I. M. S. (2008): Potential of X-Ray Spectrometry and Chemometrics to Discriminate Organic from Conventional Grown Agricultural Products. Proc. Second Scientific Conf. of International Society ofOrganic Agriculture Research (ISOFAR), Modena.

Boyko, Alex; Kovalchuk, Igor (2011): Genetic and epigenetic effects of plant-pathogen interactions: an evolutionary perspective. In *Molecular plant* 4 (6), pp. 1014–1023. DOI: 10.1093/mp/ssr022.

Bucher, Etienne; Reinders, Jon; Mirouze, Marie (2012): Epigenetic control of transposon transcription and mobility in Arabidopsis. In *Current opinion in plant biology* 15 (5), pp. 503–510. DOI: 10.1016/j.pbi.2012.08.006.

Busscher, Nicolaas; Kahl, Johannes; Andersen, Jens-Otto; Huber, Machteld; Mergardt, Gaby; Doesburg, Paul et al. (2010): Standardization of the Biocrystallization Method for Carrot Samples. In *Biological Agriculture & Horticulture* 27 (1), pp. 1–23. DOI: 10.1080/01448765.2010.10510427.

Camin, Federica; Moschella, Anna; Miselli, Francesca; Parisi, Bruno; Versini, Giuseppe; Ranalli, Paolo; Bagnaresi, Paolo (2007): Evaluation of markers for the traceability of potato tubers grown in an organicversus conventional regime. In *J. Sci. Food Agric.* 87 (7), pp. 1330–1336. DOI: 10.1002/jsfa.2853.

Capuano, Edoardo; Boerrigter-Eenling, Rita; van der Veer, Grishja; van Ruth, Saskia M. (2013): Analytical authentication of organic products: an overview of markers. In *Journal of the science of food and agriculture* 93 (1), pp. 12–28. DOI: 10.1002/jsfa.5914.

Carbonaro, Marina; Mattera, Maria; Nicoli, Stefano; Bergamo, Paolo; Cappelloni, Marsilio (2002): Modulation of Antioxidant Compounds in Organic vs Conventional Fruit (Peach, Prunus persica L., and Pear, Pyrus communis L.). In *J. Agric. Food Chem.* 50 (19), pp. 5458–5462. DOI: 10.1021/jf0202584.

Castel, Stephane E.; Martienssen, Robert A. (2013): RNA interference in the nucleus: roles for small RNAs in transcription, epigenetics and beyond. In *Nature reviews. Genetics* 14 (2), pp. 100–112. DOI: 10.1038/nrg3355.

Chalhoub, Boulos; Denoeud, France; Liu, Shengyi; Parkin, Isobel A. P.; Tang, Haibao; Wang, Xiyin et al. (2014): Plant genetics. Early allopolyploid evolution in the post-Neolithic Brassica napus oilseed genome. In *Science (New York, N.Y.)* 345 (6199), pp. 950–953. DOI: 10.1126/science.1253435.

Chan, Simon W-L; Henderson, Ian R.; Jacobsen, Steven E. (2005): Gardening the genome: DNA methylation in Arabidopsis thaliana. In *Nature reviews. Genetics* 6 (5), pp. 351–360. DOI: 10.1038/nrg1601.

Chassy, Alexander W.; Bui, Linh; Renaud, Erica N. C.; van Horn, Mark; Mitchell, Alyson E. (2006): Three-year comparison of the content of antioxidant microconstituents and several quality characteristics in organic and conventionally managed tomatoes and bell peppers. In *Journal of agricultural and food chemistry* 54 (21), pp. 8244–8252. DOI: 10.1021/jf060950p.

Chen, Pao-Yang; Cokus, Shawn J.; Pellegrini, Matteo (2010): BS Seeker: precise mapping for bisulfite sequencing. In *BMC bioinformatics* 11, p. 203. DOI: 10.1186/1471-2105-11-203.

Chen, Xiaochao; Schönberger, Brigitte; Menz, Jochen; Ludewig, Uwe (2018): Plasticity of DNA methylation and gene expression under zinc deficiency in Arabidopsis roots. In *Plant & cell physiology. DOI:* 10.1093/pcp/pcy100.

Choi, Woo-Jung; Ro, Hee-Myong; Hobbie, Erik A. (2003): Patterns of natural 15N in soils and plants from chemically and organically fertilized uplands. In *Soil Biology and Biochemistry* 35 (11), pp. 1493–1500. DOI: 10.1016/S0038-0717(03)00246-3.

Cicatelli, Angela; Baldantoni, Daniela; Iovieno, Paola; Carotenuto, Maurizio; Alfani, Anna; Feis, Italia de; Castiglione, Stefano (2014): Genetically biodiverse potato cultivars grown on a suitable agricultural soil under compost amendment or mineral fertilization. Yield, quality, genetic and epigenetic variations, soil properties. In *Science of The Total Environment* 493, pp. 1025–1035. DOI: 10.1016/j.scitotenv.2014.05.122.

Cokus, Shawn J.; Feng, Suhua; Zhang, Xiaoyu; Chen, Zugen; Merriman, Barry; Haudenschild, Christian D. et al. (2008): Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. In *Nature* 452 (7184), pp. 215–219. DOI: 10.1038/nature06745.

Condon, David E.; Tran, Phu V.; Lien, Yu-Chin; Schug, Jonathan; Georgieff, Michael K.; Simmons, Rebecca A.; Won, Kyoung-Jae (2018): Defiant: (DMRs: easy, fast, identification and ANnoTation) identifies differentially Methylated regions from iron-deficient rat hippocampus. In *BMC bioinformatics* 19 (1), p. 31. DOI: 10.1186/s12859-018-2037-1.

Cozzolino, Daniel; Holdstock, Matt; Dambergs, Robert G.; Cynkar, Wies U.; Smith, Paul A. (2009): Mid infrared spectroscopy and multivariate analysis. A tool to discriminate between organic and non-organic wines grown in Australia. In *Food Chemistry* 116 (3), pp. 761–765. DOI: 10.1016/j.foodchem.2009.03.022.

De Nadai Fernandes, Elisabete A.; Tagliaferro, Fábio S.; Azevedo-Filho, Adriano; Bode, Peter (2002): Organic coffee discrimination with INAA and data mining/KDD techniques. New perspectives for coffee trade. In *Accreditation and Quality Assurance* 7 (10), pp. 378–387. DOI: 10.1007/s00769-002-0531-6.

Deaton, Aimée M.; Bird, Adrian (2011): CpG islands and the regulation of transcription. In *Genes & development* 25 (10), pp. 1010–1022. DOI: 10.1101/gad.2037511.

del Amor, Francisco M.; Navarro, Joaquin; Aparicio, Pedro M. (2008): Isotopic discrimination as a tool for organic farming certification in sweet pepper. In *Journal of environmental quality* 37 (1), pp. 182–185. DOI: 10.2134/jeq2007.0329.

Erdmann, Robert M.; Souza, Amanda L.; Clish, Clary B.; Gehring, Mary (2014): 5-hydroxymethylcytosine is not present in appreciable quantities in Arabidopsis DNA. In *G3 (Bethesda, Md.)* 5 (1), pp. 1–8. DOI: 10.1534/g3.114.014670.

Escalona, Merly; Rocha, Sara; Posada, David (2016): A comparison of tools for the simulation of genomic next-generation sequencing data. In *Nature reviews. Genetics* 17 (8), pp. 459–469. DOI: 10.1038/nrg.2016.57.

European Commission (2008): Commission Regulation (EC) No 889/2008 of 5 September 2008 laying down detailed rules for the implementation of Council Regulation (EC) No 834/2007 on organic production and labelling of organic products with regard to organic production, labelling and control. Available online at https://eur-lex.europa.eu/eli/reg/2008/889/oj.

Finnegan, E. J.; Peacock, W. J.; Dennis, E. S. (1996): Reduced DNA methylation in Arabidopsis thaliana results in abnormal plant development. In *Proceedings of the National Academy of Sciences of the United States of America* 93 (16), pp. 8449–8454.

Flores, Kevin B.; Wolschin, Florian; Amdam, Gro V. (2013): The role of methylation of DNA in environmental adaptation. In *Integrative and comparative biology* 53 (2), pp. 359–372. DOI: 10.1093/icb/ict019.

Flusberg, Benjamin A.; Webster, Dale R.; Lee, Jessica H.; Travers, Kevin J.; Olivares, Eric C.; Clark, Tyson A. et al. (2010): Direct detection of DNA methylation during single-molecule, real-time sequencing. In *Nature methods* 7 (6), pp. 461–465. DOI: 10.1038/nmeth.1459.

Flutre, Timothée; Duprat, Elodie; Feuillet, Catherine; Quesneville, Hadi (2011): Considering transposable element diversification in de novo annotation approaches. In *PloS one* 6 (1), pp. e16526. DOI: 10.1371/journal.pone.0016526.

Frommer, M.; McDonald, L. E.; Millar, D. S.; Collis, C. M.; Watt, F.; Grigg, G. W. et al. (1992): A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. In *Proceedings of the National Academy of Sciences* 89 (5), pp. 1827–1831. DOI: 10.1073/pnas.89.5.1827.

Fujimoto, Ryo; Sasaki, Taku; Ishikawa, Ryo; Osabe, Kenji; Kawanabe, Takahiro; Dennis, Elizabeth S. (2012): Molecular mechanisms of epigenetic variation in plants. In *International journal of molecular sciences* 13 (8), pp. 9900–9922. DOI: 10.3390/ijms13089900.

García-Alcalde, Fernando; Okonechnikov, Konstantin; Carbonell, José; Cruz, Luis M.; Götz, Stefan; Tarazona, Sonia et al. (2012): Qualimap: evaluating next-generation sequencing alignment data. In *Bioinformatics (Oxford, England)* 28 (20), pp. 2678–2679. DOI: 10.1093/bioinformatics/bts503.

Ge, Yan; Wang, Lu; Li, Duanhua; Zhao, Chen; Li, Jinjun; Liu, Tao (2019): Exploring the Extended Biological Functions of the Human Copper Chaperone of Superoxide Dismutase 1. In *The protein journal* 38 (4), pp. 463–471. DOI: 10.1007/s10930-019-09824-9.

Global Next Generation Sequencing Market Assessment & Forecast: Available online: In *https://www.prnewswire.com/news-releases/global-next-generation-sequencing-market-assessment--forecast-2017---2021-300431518.html* ((accessed on 26 November 2018)).

Gosling, P.; Hodge, A.; Goodlass, G.; Bending, G. D. (2006): Arbuscular mycorrhizal fungi and organic farming. In *Agriculture, Ecosystems & Environment* 113 (1-4), pp. 17–35. DOI: 10.1016/j.agee.2005.09.009.

Grehl, Claudius; Kuhlmann, Markus; Becker, Claude; Glaser, Bruno; Grosse, Ivo (2018): How to Design a Whole-Genome Bisulfite Sequencing Experiment. In *Epigenomes* 2 (4), p. 21. DOI: 10.3390/epigenomes2040021.

Grehl, Claudius; Wagner, Marc; Lemnian, Ioana; Glaser, Bruno; Grosse, Ivo (2020): Performance of Mapping Approaches for Whole-Genome Bisulfite Sequencing Data in Crop Plants. In *Frontiers in plant science* 11, p. 176. DOI: 10.3389/fpls.2020.00176.

Guignard, Cedric; Lenouvel, Audrey; Laursen, Kristian H.; Husted, Søren (2015): Development of multi-residue methods for pesticide screening in organic food samples. In *Authentic food; core organic 2*.

Gundersen, Vagn; Bechmann, Iben Ellegaard; Behrens, Annette; Stürup, Stefan (2000): Comparative Investigation of Concentrations of Major and Trace Elements in Organic and Conventional Danish Agricultural Crops. 1. Onions (Alliumcepa Hysam) and Peas (Pisumsativum Ping Pong). In *J. Agric. Food Chem.* 48 (12), pp. 6094–6102. DOI: 10.1021/jf0009652.

Guo, Weilong; Fiziev, Petko; Yan, Weihong; Cokus, Shawn; Sun, Xueguang; Zhang, Michael Q. et al. (2013): BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data. In *BMC genomics* 14, p. 774. DOI: 10.1186/1471-2164-14-774.

Hack, H.; Gall, H.; Klemke, Th.; Klose, R.; Meier, U.; Stauss, R.; Witzenberger, A. (1993): The BBCH scale for phenological growth stages of potato (Solanum tuberosum I.). Triennale Conference of the European Association for Potato Research. Paris: INRA.

Hackett, Jamie A.; Surani, M. Azim (2013): DNA methylation dynamics during the mammalian life cycle. In *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 368 (1609), p. 20110328. DOI: 10.1098/rstb.2011.0328.

Häkkinen, Sari H.; Törrönen, A.Riitta (2000): Content of flavonols and selected phenolic acids in strawberries and Vaccinium species. Influence of cultivar, cultivation site and technique. In *Food Research International* 33 (6), pp. 517–524. DOI: 10.1016/S0963-9969(00)00086-7.

Havecker, Ericka R.; Wallbridge, Laura M.; Fedito, Paola; Hardcastle, Thomas J.; Baulcombe, David C. (2012): Metastable differentially methylated regions within Arabidopsis inbred populations are associated with modified expression of non-coding transcripts. In *PloS one* 7 (9), pp. e45242. DOI: 10.1371/journal.pone.0045242.

Hoefkens, Christine; Vandekinderen, Isabelle; Meulenaer, Bruno de; Devlieghere, Frank; Baert, Katleen; Sioen, Isabelle et al. (2009): A literature-based comparison of nutrient and contaminant contents between organic and conventional vegetables and potatoes. In *British Food Journal* 111 (10), pp. 1078–1097. DOI: 10.1108/00070700910992934.

Hoffmann, Steve; Otto, Christian; Kurtz, Stefan; Sharma, Cynthia M.; Khaitovich, Philipp; Vogel, Jörg et al. (2009): Fast mapping of short sequences with mismatches, insertions and deletions using index structures. In *PLoS computational biology* 5 (9), pp. e1000502. DOI: 10.1371/journal.pcbi.1000502.

Hossain, Md Shakhawat; Kawakatsu, Taiji; Kim, Kyung Do; Zhang, Ning; Nguyen, Cuong T.; Khan, Saad M. et al. (2017): Divergent cytosine DNA methylation patterns in single-cell, soybean root hairs. In *The New phytologist* 214 (2), pp. 808–819. DOI: 10.1111/nph.14421.

Huang, Xiaosan; Zhang, Shaoling; Li, Kongqing; Thimmapuram, Jyothi; Xie, Shaojun; Wren, Jonathan (2018): ViewBS: a powerful toolkit for visualization of high-throughput bisulfite sequencing data. In *Bioinformatics (Oxford, England)* 34 (4), pp. 708–709. DOI: 10.1093/bioinformatics/btx633.

Husted, Søren (2015): Final report for the CORE Organic II funded project - Fast methods for authentication of organic plant based foods - AuthenticFood. In *University of Copenhagen, KU-Science*.

Jeffery, E. H.; Brown, A. F.; Kurilich, A. C.; Keck, A. S.; Matusheski, N.; Klein, B. P.; Juvik, J. A. (2003): Variation in content of bioactive components in broccoli. In *Journal of Food Composition and Analysis* 16 (3), pp. 323–330. DOI: 10.1016/S0889-1575(03)00045-0.

Ji, Lexiang; Sasaki, Takahiko; Sun, Xiaoxiao; Ma, Ping; Lewis, Zachary A.; Schmitz, Robert J. (2014): Methylated DNA is over-represented in whole-genome bisulfite sequencing data. In *Frontiers in genetics* 5, p. 341. DOI: 10.3389/fgene.2014.00341.

Jühling, Frank; Kretzmer, Helene; Bernhart, Stephan H.; Otto, Christian; Stadler, Peter F.; Hoffmann, Steve (2016): metilene: fast and sensitive calling of differentially methylated regions from bisulfite sequencing data. In *Genome research* 26 (2), pp. 256–262. DOI: 10.1101/gr.196394.115.

Kahl, J.; Busscher, N.; Mergardt, G.; Mader, Paul; Dubois, D.; Ploeger, Angelika (2008): Authentication of organic wheat samples from a long-term trial using biocrystallization. Proc. Second Scientific Conf. of InternationalSociety of Organic Agriculture Research (ISOFAR), Modena,

Kazimierczak, Renata; Hallmann, Ewelina; Lipowski, Janusz; Drela, Nadzieja; Kowalik, Anna; Pussa, Tonu et al. (2014): Beetroot (Beta vulgaris L.) and naturally fermented beetroot juices from organic and conventional production: metabolomics, antioxidant levels and anticancer activity. In *Journal of the science of food and agriculture* 94 (13), pp. 2618–2629. DOI: 10.1002/jsfa.6722.

Kelly, Simon D.; Bateman, Alison S. (2010): Comparison of mineral concentrations in commercially grown organic and conventional crops – Tomatoes (Lycopersicon esculentum) and lettuces (Lactuca sativa). In *Food Chemistry* 119 (2), pp. 738–745. DOI: 10.1016/j.foodchem.2009.07.022.

Kint, Sam; Spiegelaere, Ward de; Kesel, Jonas de; Vandekerckhove, Linos; van Criekinge, Wim (2018): Evaluation of bisulfite kits for DNA methylation profiling in terms of DNA fragmentation and DNA recovery using digital PCR. In *PloS one* 13 (6), pp. e0199091. DOI: 10.1371/journal.pone.0199091.

Koch, Alexander; Joosten, Sophie C.; Feng, Zheng; Ruijter, Tim C. de; Draht, Muriel X.; Melotte, Veerle et al. (2018): Analysis of DNA methylation in cancer: location revisited. In *Nature reviews. Clinical oncology* 15 (7), pp. 459–466. DOI: 10.1038/s41571-018-0004-4.

Kohl, D. H.; Shearer, G. B.; Commoner, B. (1973): Variation of 15N in Corn and Soil Following Application of Fertilizer Nitrogen1. In *Soil Science Society of America Journal* 37 (6), p. 888. DOI: 10.2136/sssaj1973.03615995003700060028x.

Köster, Johannes; Rahmann, Sven (2012): Snakemake--a scalable bioinformatics workflow engine. In *Bioinformatics (Oxford, England)* 28 (19), pp. 2520–2522. DOI: 10.1093/bioinformatics/bts480.

Kothari, S. K.; Marschner, H.; Romheld, V. (1991): Effect of a vesicular-arbuscular mycorrhizal fungus and rhizosphere micro-organisms on manganese reduction in the rhizosphere and manganese concentrations in maize (Zea mays L.). In *New Phytol* 117 (4), pp. 649–655. DOI: 10.1111/j.1469-8137.1991.tb00969.x.

Koziol, Magdalena J.; Bradshaw, Charles R.; Allen, George E.; Costa, Ana S. H.; Frezza, Christian; Gurdon, John B. (2016): Identification of methylated deoxyadenosines in vertebrates reveals diversity in DNA modifications. In *Nature structural & molecular biology* 23 (1), pp. 24–30. DOI: 10.1038/nsmb.3145.

Krueger, Felix; Andrews, Simon R. (2011): Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. In *Bioinformatics (Oxford, England)* 27 (11), pp. 1571–1572. DOI: 10.1093/bioinformatics/btr167.

Krzywinski, Martin; Schein, Jacqueline; Birol, Inanç; Connors, Joseph; Gascoyne, Randy; Horsman, Doug et al. (2009): Circos: an information aesthetic for comparative genomics. In *Genome research* 19 (9), pp. 1639–1645. DOI: 10.1101/gr.092759.109.

Kumar, Suresh; Beena, Ananda Sankara; Awana, Monika; Singh, Archana (2017): Salt-Induced Tissue-Specific Cytosine Methylation Downregulates Expression of HKT Genes in Contrasting Wheat (Triticum aestivum L.) Genotypes. In *DNA and cell biology* 36 (4), pp. 283–294. DOI: 10.1089/dna.2016.3505.

Kunde-Ramamoorthy, Govindarajan; Coarfa, Cristian; Laritsky, Eleonora; Kessler, Noah J.; Harris, R. Alan; Xu, Mingchu et al. (2014): Comparison and quantitative verification of mapping algorithms for whole-genome bisulfite sequencing. In *Nucleic acids research* 42 (6), pp. e43. DOI: 10.1093/nar/gkt1325.

Laidig, F.; Drobek, T.; Meyer, U. (2008): Genotypic and environmental variability of yield for cultivars from 30 different crops in German official variety trials. In *Plant Breeding* 127 (6), pp. 541–547. DOI: 10.1111/j.1439-0523.2008.01564.x.

Laird, Peter W. (2010): Principles and challenges of genomewide DNA methylation analysis. In *Nature reviews. Genetics* 11 (3), pp. 191–203. DOI: 10.1038/nrg2732.

Lamesch, Philippe; Berardini, Tanya Z.; Li, Donghui; Swarbreck, David; Wilks, Christopher; Sasidharan, Rajkumar et al. (2012): The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. In *Nucleic acids research* 40 (Database issue), p. D1202-10. DOI: 10.1093/nar/gkr1090.

Langmead, Ben; Salzberg, Steven L. (2012): Fast gapped-read alignment with Bowtie 2. In *Nature methods* 9 (4), pp. 357–359. DOI: 10.1038/nmeth.1923.

Langmead, Ben; Trapnell, Cole; Pop, Mihai; Salzberg, Steven L. (2009): Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. In *Genome biology* 10 (3), pp. R25. DOI: 10.1186/gb-2009-10-3-r25.

Laursen, K. H.; Mihailova, A.; Kelly, S. D.; Epov, V. N.; Bérail, S.; Schjoerring, J. K. et al. (2013): Is it really organic? – Multi-isotopic analysis as a tool to discriminate between organic and conventional plants. In *Food Chemistry* 141 (3), pp. 2812–2820. DOI: 10.1016/j.foodchem.2013.05.068.

Laursen, K. H.; Schjoerring, J. K.; Kelly, S. D.; Husted, S. (2014): Authentication of organically grown plants – advantages and limitations of atomic spectroscopy for multi-element and stable isotope analysis. In *TrAC Trends in Analytical Chemistry* 59, pp. 73–82. DOI: 10.1016/j.trac.2014.04.008.

Laursen, Kristian H.; Schjoerring, Jan K.; Olesen, Jorgen E.; Askegaard, Margrethe; Halekoh, Ulrich; Husted, Soren (2011): Multielemental fingerprinting as a tool for authentication of organic wheat, barley, faba bean, and potato. In *Journal of agricultural and food chemistry* 59 (9), pp. 4385–4396. DOI: 10.1021/jf104928r.

Lauss, Kathrin; Wardenaar, René; Oka, Rurika; van Hulten, Marieke H A; Guryev, Victor; Keurentjes, Joost J. B. et al. (2018): Parental DNA Methylation States Are Associated with Heterosis in Epigenetic Hybrids. In *Plant physiology* 176 (2), pp. 1627–1645. DOI: 10.1104/pp.17.01054.

Law, Julie A.; Jacobsen, Steven E. (2010): Establishing, maintaining and modifying DNA methylation patterns in plants and animals. In *Nature reviews. Genetics* 11 (3), pp. 204–220. DOI: 10.1038/nrg2719.

Leggett, Richard M.; Ramirez-Gonzalez, Ricardo H.; Clavijo, Bernardo J.; Waite, Darren; Davey, Robert P. (2013): Sequencing quality assessment tools to enable data-driven informatics for high throughput genomics. In *Frontiers in genetics* 4, p. 288. DOI: 10.3389/fgene.2013.00288.

Lehesranta, Satu J.; Koistinen, Kaisa M.; Massat, Nathalie; Davies, Howard V.; Shepherd, Louise V. T.; McNicol, James W. et al. (2007): Effects of agricultural production systems and their components on protein profiles of potato tubers. In *Proteomics* 7 (4), pp. 597–604. DOI: 10.1002/pmic.200600889.

Li, Heng; Durbin, Richard (2009): Fast and accurate short read alignment with Burrows-Wheeler transform. In *Bioinformatics (Oxford, England)* 25 (14), pp. 1754–1760. DOI: 10.1093/bioinformatics/btp324.

Li, Heng; Durbin, Richard (2010): Fast and accurate long-read alignment with Burrows-Wheeler transform. In *Bioinformatics (Oxford, England)* 26 (5), pp. 589–595. DOI: 10.1093/bioinformatics/btp698.

Li, Heng; Handsaker, Bob; Wysoker, Alec; Fennell, Tim; Ruan, Jue; Homer, Nils et al. (2009): The Sequence Alignment/Map format and SAMtools. In *Bioinformatics (Oxford, England)* 25 (16), pp. 2078–2079. DOI: 10.1093/bioinformatics/btp352.

Lister, Ryan; O'Malley, Ronan C.; Tonti-Filippini, Julian; Gregory, Brian D.; Berry, Charles C.; Millar, A. Harvey; Ecker, Joseph R. (2008): Highly integrated single-base resolution maps of the epigenome in Arabidopsis. In *Cell* 133 (3), pp. 523–536. DOI: 10.1016/j.cell.2008.03.029.

Lister, Ryan; Pelizzola, Mattia; Dowen, Robert H.; Hawkins, R. David; Hon, Gary; Tonti-Filippini, Julian et al. (2009): Human DNA methylomes at base resolution show widespread epigenomic differences. In *Nature* 462 (7271), pp. 315–322. DOI: 10.1038/nature08514.

Liu, Shengyi; Snowdon, Rod; Chalhoub, Boulos (Eds.) (2018): The Brassica napus Genome. Cham: Springer International Publishing (Compendium of Plant Genomes). Available online at https://doi.org/10.1007/978-3-319-43694-4.

Lopez, Carlos M. Rodriguez; Wilkinson, Mike J. (2015): Epi-fingerprinting and epi-interventions for improved crop production and food quality. In *Frontiers in plant science* 6, p. 397. DOI: 10.3389/fpls.2015.00397.

Lu, Chungui; Hawkesford, Malcolm J.; Barraclough, Peter B.; Poulton, Paul R.; Wilson, Ian D.; Barker, Gary L.; Edwards, Keith J. (2005): Markedly different gene expression in wheat grown with organic or inorganic fertilizer. In *Proceedings. Biological sciences / The Royal Society* 272 (1575), pp. 1901–1908. DOI: 10.1098/rspb.2005.3161.

Luthria, Devanand; Singh, Ajay P.; Wilson, Ted; Vorsa, Nicholi; Banuelos, Gary S.; Vinyard, Bryan T. (2010): Influence of conventional and organic agricultural practices on the phenolic content in eggplant pulp. Plant-to-plant variation. In *Food Chemistry* 121 (2), pp. 406–411. DOI: 10.1016/j.foodchem.2009.12.055.

Lyko, F.; Ramsahoye, B. H.; Jaenisch, R. (2000): DNA methylation in Drosophila melanogaster. In *Nature* 408 (6812), pp. 538–540. DOI: 10.1038/35046205.

Mader, Paul; Fliessbach, Andreas; Dubois, David; Gunst, Lucie; Fried, Padruot; Niggli, Urs (2002): Soil fertility and biodiversity in organic farming. In *Science (New York, N.Y.)* 296 (5573), pp. 1694–1697. DOI: 10.1126/science.1071148.

Magkos, Faidon; Arvaniti, Fotini; Zampelas, Antonis (2003): Organic food: nutritious food or food for thought? A review of the evidence. In *International journal of food sciences and nutrition* 54 (5), pp. 357–371. DOI: 10.1080/09637480120092071.

Marco-Sola, Santiago; Sammeth, Michael; Guigó, Roderic; Ribeca, Paolo (2012): The GEM mapper: fast, accurate and versatile alignment by filtration. In *Nature methods* 9 (12), pp. 1185–1188. DOI: 10.1038/nmeth.2221.

Matzke, Marjori; Kanno, Tatsuo; Huettel, Bruno; Daxinger, Lucia; Matzke, Antonius J. M. (2007): Targets of RNA-directed DNA methylation. In *Current opinion in plant biology* 10 (5), pp. 512–519. DOI: 10.1016/j.pbi.2007.06.007.

Matzke, Marjori A.; Birchler, James A. (2005): RNAi-mediated pathways in the nucleus. In *Nature reviews. Genetics* 6 (1), pp. 24–35. DOI: 10.1038/nrg1500.

Mayer, Jochen; Gunst, Lucie; Mäder, Paul; Samson, Marie-Françoise; Carcea, Marina; Narducci, Valentina et al. (2015): "Productivity, quality and sustainability of winter wheat under long-term conventional and organic management in Switzerland". In *European Journal of Agronomy* 65, pp. 27–39. DOI: 10.1016/j.eja.2015.01.002.

McInroy, Gordon R.; Beraldi, Dario; Raiber, Eun-Ang; Modrzynska, Katarzyna; van Delft, Pieter; Billker, Oliver; Balasubramanian, Shankar (2016): Enhanced Methylation Analysis by Recovery of Unsequenceable Fragments. In *PloS one* 11 (3), pp. e0152322. DOI: 10.1371/journal.pone.0152322.

Mehra, Mrigaya; Gangwar, Indu; Shankar, Ravi; Houben, Andreas (2015): A Deluge of Complex Repeats. The Solanum Genome. In *PLoS ONE* 10 (8), pp. e0133962. DOI: 10.1371/journal.pone.0133962.

Mie, Axel; Laursen, Kristian Holst; Aberg, K. Magnus; Forshed, Jenny; Lindahl, Anna; Thorup-Kristensen, Kristian et al. (2014): Discrimination of conventional and organic white cabbage from a long-term field trial study using untargeted LC-MS-based metabolomics. In *Analytical and bioanalytical chemistry* 406 (12), pp. 2885–2897. DOI: 10.1007/s00216-014-7704-0.

Mitchell, Alyson E.; Hong, Yun-Jeong; Koh, Eunmi; Barrett, Diane M.; Bryant, D. E.; Denison, R. Ford; Kaffka, Stephen (2007): Ten-year comparison of the influence of organic and conventional crop management practices on the content of flavonoids in tomatoes. In *Journal of agricultural and food chemistry* 55 (15), pp. 6154–6159. DOI: 10.1021/jf070344+.

Miura, Fumihito; Enomoto, Yusuke; Dairiki, Ryo; Ito, Takashi (2012): Amplification-free whole-genome bisulfite sequencing by post-bisulfite adaptor tagging. In *Nucleic acids research* 40 (17), pp. e136. DOI: 10.1093/nar/gks454.

Muilwijk, Mirthe; Heenan, Samuel; Koot, Alex; van Ruth, Saskia M. (2015): Impact of Production Location, Production System, and Variety on the Volatile Organic Compounds Fingerprints and Sensory Characteristics of Tomatoes. In *Journal of Chemistry* 2015, pp. 1–7. DOI: 10.1155/2015/981549.

Mulero, Juana; Pardo, Francisco; Zafrilla, Pilar (2009): Effect of principal polyphenolic components in relation to antioxidant activity in conventional and organic red wines during storage. In *Eur Food Res Technol* 229 (5), pp. 807–812. DOI: 10.1007/s00217-009-1117-x.

Mulero, Juana; Pardo, Francisco; Zafrilla, Pilar (2010): Antioxidant activity and phenolic composition of organic and conventional grapes and wines. In *Journal of Food Composition and Analysis* 23 (6), pp. 569–574. DOI: 10.1016/j.jfca.2010.05.001.

Nair, Shalima S.; Luu, Phuc-Loi; Qu, Wenjia; Maddugoda, Madhavi; Huschtscha, Lily; Reddel, Roger et al. (2018): Guidelines for whole genome bisulphite sequencing of intact and FFPET DNA on the Illumina HiSeq X Ten. In *Epigenetics & chromatin* 11 (1), p. 24. DOI: 10.1186/s13072-018-0194-0.

Niederhuth, Chad E.; Schmitz, Robert J. (2014): Covering your bases: inheritance of DNA methylation in plant genomes. In *Molecular plant* 7 (3), pp. 472–480. DOI: 10.1093/mp/sst165.

Oehl, Fritz; Sieverding, Ewald; Mader, Paul; Dubois, David; Ineichen, Kurt; Boller, Thomas; Wiemken, Andres (2004): Impact of long-term conventional and organic farming on the diversity of arbuscular mycorrhizal fungi. In *Oecologia* 138 (4), pp. 574–583. DOI: 10.1007/s00442-003-1458-2.

Olova, Nelly; Krueger, Felix; Andrews, Simon; Oxley, David; Berrens, Rebecca V.; Branco, Miguel R.; Reik, Wolf (2018): Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting DNA methylation data. In *Genome biology* 19 (1), p. 33. DOI: 10.1186/s13059-018-1408-2.

Omony, Jimmy; Nussbaumer, Thomas; Gutzat, Ruben (2019): DNA methylation analysis in plants. Review of computational tools and future perspectives. In *Briefings in bioinformatics* 38 (5), p. 285. DOI: 10.1093/bib/bbz039.

Ordóñez-Santos, Luis Eduardo; Vázquez-Odériz, M. Lourdes; Romero-Rodríguez, M. Angeles (2011): Micronutrient contents in organic and conventional tomatoes (Solanum lycopersicum L.). In *International Journal of Food Science & Technology* 46 (8), pp. 1561–1568. DOI: 10.1111/j.1365-2621.2011.02648.x.

Otto, Christian; Stadler, Peter F.; Hoffmann, Steve (2012): Fast and sensitive mapping of bisulfite-treated sequencing data. In *Bioinformatics (Oxford, England)* 28 (13), pp. 1698–1704. DOI: 10.1093/bioinformatics/bts254.

Paolini, Mauro; Ziller, Luca; Laursen, Kristian Holst; Husted, Soren; Camin, Federica (2015): Compound-Specific delta(1)(5)N and delta(1)(3)C Analyses of Amino Acids for Potential Discrimination between Organically and Conventionally Grown Wheat. In *Journal of agricultural and food chemistry* 63 (25), pp. 5841–5850. DOI: 10.1021/acs.jafc.5b00662.

Peat, Julian R.; Smallwood, Sébastien A. (2018): Low Input Whole-Genome Bisulfite Sequencing Using a Post-Bisulfite Adapter Tagging Approach. In *Methods in molecular biology (Clifton, N.J.)* 1708, pp. 161–169. DOI: 10.1007/978-1-4939-7481-8_9.

Pedersen, B.; Eyring, K.; De, S.; Yang, I.; Schwartz, D. (2014): Fast and accurate alignment of long bisulfite-seq reads. Prepring: arXiv:1401.1129v2. In *Bioinformatics (Oxford, England)*, 5/13/2014.

Putzig, Curtis L.; Leugers, M. Anne; McKelvy, Marianne L.; Mitchell, Gary E.; Nyquist, Richard A. (1994): Infrared Spectroscopy. In *Anal. Chem.* (66), pp. 26–66.

Rackham, Owen J. L.; Dellaportas, Petros; Petretto, Enrico; Bottolo, Leonardo (2015): WGBSSuite: simulating whole-genome bisulphite sequencing data and benchmarking differential DNA methylation analysis tools. In *Bioinformatics (Oxford, England)* 31 (14), pp. 2371–2373. DOI: 10.1093/bioinformatics/btv114.

Raigon, Maria D.; Rodriguez-Burruezo, Adrian; Prohens, Jaime (2010): Effects of organic and conventional cultivation methods on composition of eggplant fruits. In *Journal of agricultural and food chemistry* 58 (11), pp. 6833–6840. DOI: 10.1021/jf904438n.

Raine, Amanda; Liljedahl, Ulrika; Nordlund, Jessica (2018): Data quality of whole genome bisulfite sequencing on Illumina platforms. In *PloS one* 13 (4), pp. e0195972. DOI: 10.1371/journal.pone.0195972.

Rapisarda, Paolo; Calabretta, Maria Luisa; Romano, Gabriella; Intrigliolo, Francesco (2005): Nitrogen Metabolism Components as a Tool To Discriminate between Organic and Conventional Citrus Fruits. In *J. Agric. Food Chem.* 53 (7), pp. 2664–2669. DOI: 10.1021/jf048733g.

Remy, Estelle; Duque, Paula (2014): Beyond cellular detoxification: a plethora of physiological roles for MDR transporter homologs in plants. In *Frontiers in physiology* 5, p. 201. DOI: 10.3389/fphys.2014.00201.

Ren, Huifeng; Endo, Hideaki; Hayashi, Tetsuhito (2001): Antioxidative and antimutagenic activities and polyphenol content of pesticide-free and organically cultivated green vegetables using water-soluble chitosan as a soil modifier and leaf surface spray. In *J. Sci. Food Agric.* 81 (15), pp. 1426–1432. DOI: 10.1002/jsfa.955.abs.

Robinson, James T.; Thorvaldsdóttir, Helga; Winckler, Wendy; Guttman, Mitchell; Lander, Eric S.; Getz, Gad; Mesirov, Jill P. (2011): Integrative genomics viewer. In *Nature biotechnology* 29 (1), pp. 24–26. DOI: 10.1038/nbt.1754.

Rogers, Karyne M. (2008): Nitrogen isotopes as a screening tool to determine the growing regimen of some organic and nonorganic supermarket produce from New Zealand. In *Journal of agricultural and food chemistry* 56 (11), pp. 4078–4083. DOI: 10.1021/jf800797w.

Rosen, Carl J.; Allan, Deborah L. (2007): Exploring the benefits of organic nutrient sources for crop production and soil quality. In *HortTechnology* 17 (4), pp. 422–430.

Rossetto, M.R.M.; Vianello, F.; Rocha, S. A.; Lima, G.P.P. (2009): Antioxidant substances and pesticide in parts of beet organic and conventional manure. In *African Journal of Plant Science* 3, pp. 245–253.

Sánchez, María-Teresa; Garrido-Varo, Ana; Guerrero, José-Emilio; Pérez-Marín, Dolores (2013): NIRS technology for fast authentication of green asparagus grown under organic and conventional production systems. In *Postharvest Biology and Technology* 85, pp. 116–123. DOI: 10.1016/j.postharvbio.2013.05.008.

Santelia, Diana; Vincenzetti, Vincent; Azzarello, Elisa; Bovet, Lucien; Fukao, Yoichiro; Düchtig, Petra et al. (2005): MDR-like ABC transporter AtPGP4 is involved in auxin-mediated lateral root and root hair development. In *FEBS letters* 579 (24), pp. 5399–5406. DOI: 10.1016/j.febslet.2005.08.061.

Schmidt, Hanns-ludwig; Roßmann, Andreas; Voerkelius, Susanne; Schnitzler, Wilfried H.; Georgi, Michael; Graßmann, Johanna et al. (2005): Isotope characteristics of vegetables and wheat from conventional and organic production. In *Isotopes in environmental and health studies* 41 (3), pp. 223–228. DOI: 10.1080/10256010500230072.

Schmutz, Jeremy; Cannon, Steven B.; Schlueter, Jessica; Ma, Jianxin; Mitros, Therese; Nelson, William et al. (2010): Genome sequence of the palaeopolyploid soybean. In *Nature* 463 (7278), pp. 178–183. DOI: 10.1038/nature08670.

Schnable, Patrick S.; Ware, Doreen; Fulton, Robert S.; Stein, Joshua C.; Wei, Fusheng; Pasternak, Shiran et al. (2009): The B73 maize genome: complexity, diversity, and dynamics. In *Science (New York, N.Y.)* 326 (5956), pp. 1112–1115. DOI: 10.1126/science.1178534.

Selker, E. U.; Stevens, J. N. (1985): DNA methylation at asymmetric sites is associated with numerous transition mutations. In *Proceedings of the National Academy of Sciences of the United States of America* 82 (23), pp. 8114–8118.

Shafi, Adib; Mitrea, Cristina; Nguyen, Tin; Draghici, Sorin (2017): A survey of the approaches for identifying differential methylation using bisulfite sequencing data. In *Briefings in bioinformatics. DOI:* 10.1093/bib/bbx013.

Shen, Huaishun; He, Hang; Li, Jigang; Chen, Wei; Wang, Xuncheng; Guo, Lan et al. (2012): Genome-wide analysis of DNA methylation and gene expression changes in two Arabidopsis ecotypes and their reciprocal hybrids. In *The Plant cell* 24 (3), pp. 875–892. DOI: 10.1105/tpc.111.094870.

Sherman-Broyles, Sue; Bombarely, Aureliano; Grimwood, Jane; Schmutz, Jeremy; Doyle, Jeff (2014): Complete plastome sequences from Glycine syndetika and six additional perennial wild relatives of soybean. In *G3 (Bethesda, Md.)* 4 (10), pp. 2023–2033. DOI: 10.1534/g3.114.012690.

Shi, Dong-Qiao; Ali, Iftikhar; Tang, Jun; Yang, Wei-Cai (2017): New Insights into 5hmC DNA Modification: Generation, Distribution and Function. In *Frontiers in genetics* 8, p. 100. DOI: 10.3389/fgene.2017.00100.

Shimada, Tomoo; Watanabe, Etsuko; Tamura, Kentaro; Hayashi, Yasuko; Nishimura, Mikio; Hara-Nishimura, Ikuko (2002): A vacuolar sorting receptor PV72 on the membrane of vesicles that accumulate precursors of seed storage proteins (PAC vesicles). In *Plant & cell physiology* 43 (10), pp. 1086–1095. DOI: 10.1093/pcp/pcf152.

Simpson, Jared T.; Workman, Rachael E.; Zuzarte, P. C.; David, Matei; Dursi, L. J.; Timp, Winston (2017): Detecting DNA cytosine methylation using nanopore sequencing. In *Nature methods* 14 (4), pp. 407–410. DOI: 10.1038/nmeth.4184.

Stöger, Reinhard; Scaife, Paula J.; Shephard, Freya; Chakrabarti, Lisa (2017): Elevated 5hmC levels characterize DNA of the cerebellum in Parkinson's disease. In *NPJ Parkinson's disease* 3, p. 6. DOI: 10.1038/s41531-017-0007-3.

Sturm, Martina; Lojen, Sonja (2011): Nitrogen isotopic signature of vegetables from the Slovenian market and its suitability as an indicator of organic production. In *Isotopes in environmental and health studies* 47 (2), pp. 214–220. DOI: 10.1080/10256016.2011.570865.

Sun, Xiwei; Han, Yi; Zhou, Liyuan; Chen, Enguo; Lu, Bingjian; Liu, Yong et al. (2018): A comprehensive evaluation of alignment software for reduced representation bisulfite sequencing data. In *Bioinformatics (Oxford, England)* 34 (16), pp. 2715–2723. DOI: 10.1093/bioinformatics/bty174.

Sun, Zhifu; Cunningham, Julie; Slager, Susan; Kocher, Jean-Pierre (2015): Base resolution methylome profiling: considerations in platform selection, data preprocessing and analysis. In *Epigenomics* 7 (5), pp. 813–828. DOI: 10.2217/epi.15.21.

Surdonja, Korana; Eggert, Kai; Hajirezaei, Mohammad-Reza; Harshavardhan, Vokkaliga; Seiler, Christiane; Wirén, Nicolaus von et al. (2017): Increase of DNA Methylation at the HvCKX2.1 Promoter by Terminal Drought Stress in Barley. In *Epigenomes* 1 (2), p. 9. DOI: 10.3390/epigenomes1020009.

Szulc, Małgorzata; Kahl, Johannes; Busscher, Nicolaas; Mergardt, Gaby; Doesburg, Paul; Ploeger, Angelika (2010): Discrimination between organically and conventionally grown winter wheat farm pair samples using the copper chloride crystallisation method in combination with computerised image analysis. In *Computers and Electronics in Agriculture* 74 (2), pp. 218–222. DOI: 10.1016/j.compag.2010.08.001.

Tanaka, Kazuo; Okamoto, Akimitsu (2007): Degradation of DNA by bisulfite treatment. In *Bioorganic & medicinal chemistry letters* 17 (7), pp. 1912–1915. DOI: 10.1016/j.bmcl.2007.01.040.

Tarasov, Artem; Vilella, Albert J.; Cuppen, Edwin; Nijman, Isaac J.; Prins, Pjotr (2015): Sambamba: fast processing of NGS alignment formats. In *Bioinformatics (Oxford, England)* 31 (12), pp. 2032–2034. DOI: 10.1093/bioinformatics/btv098.

Tran, Hong; Porter, Jacob; Sun, Ming-An; Xie, Hehuang; Zhang, Liqing (2014): Objective and comprehensive evaluation of bisulfite short read mapping tools. In *Advances in bioinformatics* 2014, p. 472045. DOI: 10.1155/2014/472045.

Tsuji, Junko; Weng, Zhiping (2016): Evaluation of preprocessing, mapping and postprocessing algorithms for analyzing whole genome bisulfite sequencing data. In *Briefings in bioinformatics* 17 (6), pp. 938–952. DOI: 10.1093/bib/bbv103.

Vallverdu-Queralt, Anna; Medina-Remon, Alexander; Casals-Ribes, Isidre; Amat, Mercedes; Lamuela-Raventos, Rosa Maria (2011): A metabolomic approach differentiates between conventional and organic ketchups. In *Journal of agricultural and food chemistry* 59 (21), pp. 11703–11710. DOI: 10.1021/jf202822s.

van Bel, Michiel; Diels, Tim; Vancaester, Emmelien; Kreft, Lukasz; Botzki, Alexander; van de Peer, Yves et al. (2018): PLAZA 4.0: an integrative resource for functional, evolutionary and comparative plant genomics. In *Nucleic acids research* 46 (D1), pp. D1190-D1196. DOI: 10.1093/nar/gkx1002.

van Dijk, Jeroen P.; Cankar, Katarina; Hendriksen, Peter J. M.; Beenen, Henriek G.; Zhu, Ming; Scheffer, Stanley et al. (2012): The Identification and Interpretation of Differences in the Transcriptomes of Organically and Conventionally Grown Potato Tubers. In *J. Agric. Food Chem.* 60 (9), pp. 2090–2101. DOI: 10.1021/jf204696w.

van Gurp, Thomas P.; Wagemaker, Niels C A M; Wouters, Bjorn; Vergeer, Philippine; Ouborg, Joop N. J.; Verhoeven, Koen J. F. (2016): epiGBS: reference-free reduced representation bisulfite sequencing. In *Nature methods* 13 (4), pp. 322–324. DOI: 10.1038/nmeth.3763.

Verhoeven, Koen J. F.; Jansen, Jeroen J.; van Dijk, Peter J.; Biere, Arjen (2010): Stress-induced DNA methylation changes and their heritability in asexual dandelions. In *The New phytologist* 185 (4), pp. 1108–1118. DOI: 10.1111/j.1469-8137.2009.03121.x.

Vrček, Ivana Vinković; Bojić, Mirza; Žuntar, Irena; Mendaš, Gordana; Medić-Šarić, Marica (2011): Phenol content, antioxidant activity and metal composition of Croatian wines deriving from organically and conventionally grown grapes. In *Food Chemistry* 124 (1), pp. 354–361. DOI: 10.1016/j.foodchem.2010.05.118.

Wang, Lin; Xie, Jiahui; Hu, Jiantuan; Lan, Binyuan; You, Chenjiang; Li, Fenglan et al. (2018): Comparative epigenomics reveals evolution of duplicated genes in potato and tomato. In *The Plant journal : for cell and molecular biology* 93 (3), pp. 460–471. DOI: 10.1111/tpj.13790.

Wang, Ming-Bo; Masuta, Chikara; Smith, Neil A.; Shimura, Hanako (2012): RNA silencing and plant viral diseases. In *Molecular plant-microbe interactions : MPMI* 25 (10), pp. 1275–1285. DOI: 10.1094/MPMI-04-12-0093-CR.

Wang, Shiow Y.; Chen, Chi-Tsun; Sciarappa, William; Wang, Chien Y.; Camp, Mary J. (2008): Fruit quality, antioxidant capacity, and flavonoid content of organically and conventionally grown blueberries. In *Journal of agricultural and food chemistry* 56 (14), pp. 5788–5794. DOI: 10.1021/jf703775r.

Wang, Xi-liang; Song, Shu-hui; Wu, Yong-Sheng; Li, Yu-Li; Chen, Ting-ting; Huang, Zhi-yuan et al. (2015): Genome-wide mapping of 5-hydroxymethylcytosine in three rice cultivars reveals its preferential localization in transcriptionally silent transposable element genes. In *Journal of experimental botany* 66 (21), pp. 6651–6663. DOI: 10.1093/jxb/erv372.

Warnecke, Peter M.; Stirzaker, Clare; Song, Jenny; Grunau, Christoph; Melki, John R.; Clark, Susan J. (2002): Identification and resolution of artifacts in bisulfite sequencing. In *Methods* 27 (2), pp. 101–107. DOI: 10.1016/S1046-2023(02)00060-9.

Wassenegger, Michael; Heimes, Sabine; Riedel, Leonhard; Sänger, Heinz L. (1994): RNA-directed de novo methylation of genomic sequences in plants. In *Cell* 76 (3), pp. 567–576. DOI: 10.1016/0092-8674(94)90119-8.

Watanabe, Etsuko; Shimada, Tomoo; Tamura, Kentaro; Matsushima, Ryo; Koumoto, Yasuko; Nishimura, Mikio; Hara-Nishimura, Ikuko (2004): An ER-localized form of PV72, a seed-specific vacuolar sorting receptor, interferes the transport of an NPIR-containing proteinase in Arabidopsis leaves. In *Plant & cell physiology* 45 (1), pp. 9–17. DOI: 10.1093/pcp/pch012.

Wolf, Paul G.; Sessa, Emily B.; Marchant, Daniel Blaine; Li, Fay-Wei; Rothfels, Carl J.; Sigel, Erin M. et al. (2015): An Exploration into Fern Genome Space. In *Genome biology and evolution* 7 (9), pp. 2533–2544. DOI: 10.1093/gbe/evv163.

Wreczycka, Katarzyna; Gosdschan, Alexander; Yusuf, Dilmurat; Grüning, Björn; Assenov, Yassen; Akalin, Altuna (2017): Strategies for analyzing bisulfite sequencing data. In *Journal of biotechnology* 261, pp. 105–115. DOI: 10.1016/j.jbiotec.2017.08.007.

Wu, Thomas D.; Nacu, Serban (2010): Fast and SNP-tolerant detection of complex variants and splicing in short reads. In *Bioinformatics (Oxford, England)* 26 (7), pp. 873–881. DOI: 10.1093/bioinformatics/btq057.

Wu, Thomas D.; Reeder, Jens; Lawrence, Michael; Becker, Gabe; Brauer, Matthew J. (2016): GMAP and GSNAP for Genomic Sequence Alignment: Enhancements to Speed, Accuracy, and Functionality. In *Methods in molecular biology (Clifton, N.J.)* 1418, pp. 283–334. DOI: 10.1007/978-1-4939-3578-9_15.

Wyatt, G. R.; Cohen, S. S. (1953): The bases of the nucleic acids of some bacterial and animal viruses. The occurrence of 5-hydroxymethylcytosine. In *Biochem. J.* 55 (5), pp. 774–782. DOI: 10.1042/bj0550774.

Xi, Yuanxin; Li, Wei (2009): BSMAP: whole genome bisulfite sequence MAPping program. In *BMC bioinformatics* 10, p. 232. DOI: 10.1186/1471-2105-10-232.

Xie, Min; Chung, Claire Yik-Lok; Li, Man-Wah; Wong, Fuk-Ling; Wang, Xin; Liu, Ailin et al. (2019): A reference-grade wild soybean genome. In *Nature communications* 10 (1), p. 1216. DOI: 10.1038/s41467-019-09142-9.

Xu, Xun; Pan, Shengkai; Cheng, Shifeng; Zhang, Bo; Mu, Desheng; Ni, Peixiang et al. (2011): Genome sequence and analysis of the tuber crop potato. In *Nature* 475 (7355), pp. 189–195. DOI: 10.1038/nature10158.

Zakrzewski, Falk; Schmidt, Martin; van Lijsebettens, Mieke; Schmidt, Thomas (2017): DNA methylation of retrotransposons, DNA transposons and genes in sugar beet (Beta vulgaris L.). In *The Plant journal : for cell and molecular biology* 90 (6), pp. 1156–1175. DOI: 10.1111/tpj.13526.

Zhang, Huiming; Lang, Zhaobo; Zhu, Jian-Kang (2018): Dynamics and function of DNA methylation in plants. In *Nature reviews. Molecular cell biology* 19 (8), pp. 489–506. DOI: 10.1038/s41580-018-0016-z.

Zhou, Jia; Sears, Renee L.; Xing, Xiaoyun; Zhang, Bo; Li, Daofeng; Rockweiler, Nicole B. et al. (2017): Tissue-specific DNA methylation is conserved across human, mouse, and rat, and driven by primary sequence conservation. In *BMC genomics* 18 (1), p. 724. DOI: 10.1186/s12864-017-4115-6.

Ziller, Michael J.; Hansen, Kasper D.; Meissner, Alexander; Aryee, Martin J. (2015): Coverage recommendations for methylation analysis by whole-genome bisulfite sequencing. In *Nature methods* 12 (3), p. 230-2, 1 p following 232. DOI: 10.1038/nmeth.3152.

Zörb, Christian; Betsche, Thomas; Langenkämper, Georg (2009a): Search for Diagnostic Proteins To Prove Authenticity of Organic Wheat Grains (Triticum aestivum L.). In *J. Agric. Food Chem.* 57 (7), pp. 2932–2937. DOI: 10.1021/jf802923r.

Zörb, Christian; Niehaus, Karsten; Barsch, Aiko; Betsche, Thomas; Langenkamper, Georg (2009b): Levels of compounds and metabolites in wheat ears and grains in organic and conventional agriculture. In *Journal of agricultural and food chemistry* 57 (20), pp. 9555–9562. DOI: 10.1021/jf9019739.

8. Authors' contributions to the included manuscripts and publications

In this dissertation in total three manuscripts or publications are included. The estimated contribution of all authors should be listed here:

How to Design a Whole-Genome Bisulfite Sequencing Experiment (DOI: 10.3390/epigenomes2040021)

Claudius Grehl	55 %	conceptualization, methodology, investigation, writing—original draft preparation, visualization, project administration, funding acquisition
Markus Kuhlmann	15 %	conceptualization, methodology, investigation, writing—original draft preparation, writing—review and editing
Claude Becker	5 %	methodology, investigation, writing—review and editing
Bruno Glaser	15 %	writing—review and editing, supervision, project administration, funding acquisition
Ivo Grosse	10 %	supervision, project administration, funding acquisition

Performance of Mapping Approaches for Whole-Genome Bisulfite Sequencing Data in Crop Plants (DOI: 10.3389/fpls.2020.00176)

Claudius Grehl	60 %	conceptualization, methodology, investigation, writing—original draft preparation, visualization, project administration, funding acquisition
Marc Wagner	10 %	conceptualization, methodology, investigation, visualization
Ioana Lemnian	5 %	writing—original draft preparation, writing—review and editing
Bruno Glaser	10 %	writing—review and editing, supervision, project administration, funding acquisition
Ivo Grosse	15 %	writing—review and editing, supervision, project administration, funding acquisition

Differences in DNA methylation patterns of organic and conventionally grown potato (Solanum tuberosum) samples (Manuscript, submitted soon/under review)

Claudius Grehl	45 %	conceptualization, methodology, investigation, writing—original draft preparation, visualization, project administration, funding acquisition
Markus Kuhlmann	15 %	conceptualization, methodology, investigation, writing—original draft preparation, writing—review and editing
Paul Mäder	5 %	writing—review and editing, sample preparation
Jochen Mayer	5 %	writing—review and editing, sample preparation
Lothar Altschmied	5 %	conceptualization, writing—review and editing
Ivo Grosse	10 %	conceptualization, writing—review and editing, supervision, project administration, funding acquisition
Bruno Glaser	15 %	conceptualization, writing—review and editing, supervision, project administration, funding acquisition

9. Curriculum vitae

CLAUDIUS GREHL (M. SC.)

EDUCATION

01.10.2014 – 30.09.2017 Martin-Luther-University Halle-Wittenberg
Halle (Saale) Master of Science, Agriculture

01.10.2011 – 30.09.2014 Leipzig University
Leipzig Bachelor of Science, Biology

PROFESSIONAL EXPERIENCE

01.04.2021 – today Institute for Energy and Environmental Research (ifeu) gGmbH , Heidelberg

Data Scientist and Life cycle assessor

+ Life cycle assessments (i. a. land use, water footprint)
+ Biomass flow and potential analysis
+ Food and renewable raw materials

01.10.2017 – 30.09.2020 Martin-Luther-University Halle-Wittenberg Halle (Saale)

Data Scientist and federal scholarship holder (PhD promotion)

+ Institute of agricultural and nutritional sciences,
  Soil-biogeochemistry group (Prof. Dr. Bruno Glaser)
+ Institute of Computer Science,
  Bioinformatics group (Prof. Dr. Ivo Grosse)
+ project specific data analytics, visualization and discussion
  (e.g. for medicinal, soil chemistry, biochemistry or genetics datasets)
+ project management, planning and communication
+ publication writing and acquisition of third party funds

01.05.2018 – 31.07.2018 Leibniz-Institute for Plant Genetics and Crop Plant Research (IPK), Gatersleben

Computational Biologist

+ Heterosis group
+ Research areas: data science, epigenetics and
  next-generation-sequencing data analytics

01.10.2013 – 31.12.2013 Kew, Royal Botanic Gardens London (UK)

Research Assistant

+ Department of Conservation Biotechnology, Jodrell Laboratory

SKILLS

Languages

German	mother tongue
English	very well (C1)
Italian	basic
French	basic

Programming languages & IT

R	well
Python	advanced
Bash/Unix/Linux	advanced

Project management, moderation and teaching (selection)

- University didactics certificate (2020)
- Research and problem-oriented learning & teaching (2019)
- University teaching I&II (2019)
- Moderation of learning processes in science & teaching (2019)
- Agile project management (2018)
- Introduction to economics (2014)

AWARDS

- Graduation Stipendship (federal, Saxony-Anhalt) 10/2017 – 9/2020

  PhD promotion on the topic: DNA methylation patterns of organic and conventional food samples
- Scholarship of the International Building Exhibition (IBA) 6/2017

  Interdisciplinary project work on the topic: StadtLand – 1.500ha future, landscape typologies of the 21st century

REFERENCES

Prof. Dr. Bruno Glaser, Professor for Soil-Biogeochemistry, Universität Halle-Wittenberg, Von-Seckendorff-Platz 3, 06120 Halle (Saale), bruno.glaser@landw.uni-halle.de, Tel.: +49 (0)345-5522532

Prof. Dr. Ivo Grosse, Professor for Bioinformatics, University Halle-Wittenberg, Von-Seckendorff-Platz 1, 06120 Halle (Saale), ivo.grosse@informatik.uni-halle.de, Tel.: +49 (0)345-5524774

10. List of publications

+ **Grehl C.**, Kuhlmann M., Mäder P., Mayer J., Altschmied L., Grosse I., Glaser B., (2021 expected) Differences in DNA methylation patterns of organic and conventionally grown potato (Solanum tuberosum) samples (Manuscript, submitted soon/under review)

+ **Grehl C.**, Schultheiß C., Hoffmann K., Binder M., Altmann T., Grosse I., Kuhlmann M., (2021 expected) Detection of SARS-CoV-2 derived small RNAs and changes in circulating small RNAs associated with COVID-19 (Manuscript, submitted soon/under review)

+ **Grehl C.**, Wagner M., Lemnian I., Glaser B., Grosse I., (2020) Performance of Mapping Approaches for Whole-Genome Bisulfite Sequencing Data in Crop Plants, Frontiers in Plant Science, Special Issue: Advances in Applied Bioinformatics in Crops, DOI: 10.3389/fpls.2020.00176

+ Lemma B., **Grehl C.**, Zech M., Mekonnen B., Zech W., Nemomissa S., Bekele T. and Glaser B. (2019) Phenolic Compounds as Unambiguous Chemical Markers for the Identification of Keystone Plant Species in the Bale Mountains, Ethiopia. Plants 8(7), 228, DOI: 10.3390/plants8070228

+ **Grehl C.**, Kuhlmann M., Becker C., Glaser B., Grosse I. (2018) How to Design a Whole-Genome Bisulfite Sequencing Experiment. Epigenomes 2018, 2(4), 21, DOI: 10.3390/epigenomes2040021

+ **Grehl C.**, Tscherteu B., Geier V., Steverding M. (2018) Sprechen wir über die Landschaft – Eine Aufforderung, die Perspektiven zu erweitern. Zoll+ laut, Österreichische Schriftenreihe für Landschaft und Freiraum, 28(33)

11. Conference contributions

+ Applied Bioinformatics for Crops, Gatersleben Research Conference (Germany) 18.-20. march 2019, talk: EpiOrg: DNA methylation patterns of organic and conventional food samples

+ Plant Epigenetics Conference, Angers (France) 29.-31. october 2018, poster: EpiOrg: Detecting food fraud by means of DNA methylation

+ Statistical Data Analysis for Genome Scale Biology, Brixen (Italy), CSAMA 8.-13. july 2018, talk: DNA methylation patterns of organic and conventional food samples

12. Declaration under Oath / Eidesstattliche Erklärung

I declare under penalty of perjury that this thesis is my own work entirely and has been written without any help from other people. I used only the sources mentioned and included all the citations correctly both in word or content.

Ich erkläre an Eides statt, dass ich die Arbeit selbstständig und ohne fremde Hilfe verfasst, keine anderen als die von mir angegebenen Quellen und Hilfsmittel benutzt und die den benutzten Werken wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

____

Claudius Grehl

13. Declaration concerning criminal record and pending investigations / Erklärung über bestehende Vorstrafen und anhängige Ermittlungsverfahren

I hereby declare, that I have no criminal record and there are no preliminary criminal proceedings pending against me.

Hiermit erkläre ich, dass ich weder vorbestraft bin noch dass gegen mich Ermittlungsverfahren anhängig sind.

____

Claudius Grehl

14. Acknowledgments

This dissertation has been accomplished and facilitated with and by several people. Please apologize that I refrain from mentioning their individual contributions as this would remain incomplete.

I want to thank:

Prof. Dr. Bruno Glaser

Prof. Dr. Ivo Grosse

Dr. Markus Kuhlmann

The entire former Bioinformatics Group at Halle-University (Samar Fatma, Dr. Alexander Gabel, Dr. Jan Grau, Silvio Weging, Claus Weinholdt, Ioana Lemnian and Martin Porsch)

Tobias Bromm, Aaron Tinschert, Dr. Bruk Lemma and all members of the Soil Biogeochemistry Group.

The employees form ITZ.

VW Foundation and the Federal State of Saxony-Anhalt (Graduation Stipendship), as well as the funding programme Open Access Publishing by the German Research Foundation (DFG) for financial support.

Ute Gradmann and Jannis Grehl

Jonathan Schörnig

Constantin Schörnig

Ansgar Oltmanns

Jan Schütter

Maximilian Rembe

All members of Hohe4.

All Neckar Crocodiles.

My family (Elke and Stefan Grehl, Luisa Wolf, Alexander Grehl, Adelheid and Franz Grehl, Magdalena and Norbert Dierich).

All I forgot to mention.