Don't be scared of the genome's 5th base – Explaining phenotypic variability and evolutionary dynamics through DNA methylation analysis

Joerg Tost[1]

[1]CEA-Institut de Biologie François Jacob, Université Paris-Saclay

May 17, 2023

Abstract

Epigenetic processes have taken center stage for the investigation of many biological processes and epigenetic modifications have shown to influence phenotype, morphology and behavioral traits such as stress resistance by affecting gene regulation and expression without altering the underlying genomic sequence. The multiple molecular layers of epigenetics synergistically construct the cell type-specific gene regulatory networks. DNA methylation occurring on the 5' carbon of cytosines in different genomic sequence contexts is the most studied epigenetic modification. DNA methylation has been shown to provide a molecular record of a large variety of environmental factors, which might be persistent through the entire lifetime of an organisms and even be passed onto the offspring. Animals might display altered phenotypes mediated by epigenetic modifications depending on the developmental stage or the environmental conditions as well as during evolution. Therefore, the analysis of DNA methylation patterns might allow deciphering previous exposures, explaining ecologically relevant phenotypic diversity and predicting evolutionary trajectories enabling accelerated adaption to changing environmental conditions. Despite the explanatory potential of DNA methylation. studies of DNA methylation are still scarce in the field of ecology. This might be at least partly due to the complexity of DNA methylation analysis and the interpretation of the acquired data. In the current issue of Molecular Ecology Resources, Laine and colleagues (2023) provide a detailed summary of guidelines and valuable recommendations for researchers in the field of ecology to avoid common pitfalls and perform interpretable genome-wide DNA methylation analyses.

Don't be scared of the genome's 5th base –

Explaining phenotypic variability and evolutionary dynamics through DNA methylation analysis

Jörg Tost*

Laboratory for Epigenetics and Environment, Centre National de Recherche en Génomique Humaine, CEA-Institut de Biologie François Jacob, Université Paris-Saclay, 2 rue Gaston Crémieux, 91000 Evry, France

*Corresponding author, E-mail: tost@cnrgh.fr

Epigenetic processes have taken center stage for the investigation of many biological processes and epigenetic modifications have shown to influence phenotype, morphology and behavioral traits such as stress resistance by affecting gene regulation and expression without altering the underlying genomic sequence. The multiple molecular layers of epigenetics synergistically construct the cell type-specific gene regulatory networks, characterized by a high degree of plasticity and redundancy to create cell-type specific morphology and function. DNA methylation occurring on the 5' carbon of cytosines in different genomic sequence contexts is the most studied epigenetic modification. DNA methylation has been shown to provide a molecular record of a large variety of environmental factors, which might be persistent through the entire

lifetime of an organisms and even be passed onto the offspring. Animals might display altered phenotypes mediated by epigenetic modifications depending on the developmental stage or the environmental conditions as well as during evolution. Therefore, the analysis of DNA methylation patterns might allow deciphering previous exposures, explaining ecologically relevant phenotypic diversity and predicting evolutionary trajectories enabling accelerated adaption to changing environmental conditions. Despite the explanatory potential of DNA methylation integrating genetic and environmental factors to shape phenotypic variation and contribute significantly to evolutionary dynamics, studies of DNA methylation are still scarce in the field of ecology. This might be at least partly due to the complexity of DNA methylation analysis and the interpretation of the acquired data. In the current issue of Molecular Ecology Resources, Laine and colleagues (2023) provide a detailed summary of guidelines and valuable recommendations for researchers in the field of ecology to avoid common pitfalls and perform interpretable genome-wide DNA methylation analyses.

A large variety of techniques to study DNA methylation at high-resolution either genome-wide or at specific loci have been devised and this field is rapidly and constantly evolving with new methods being developed due to technological advances as well as our increasing knowledge on the gene regulatory landscape created by epigenetic modifications (Tost, 2022). Epigenome-wide association patterns analyzing genome-wide DNA methylation patterns have been performed in many human cohorts demonstrating altered DNA methylation associated with most phenotypes that have so far been investigated as well as in function of exposure to environmental conditions ranging from temperature to chemical pollutants providing a solid scientific basis for the need of similar studies in non-model animals.

Laine and colleagues focus on bisulfite sequencing, which is the most commonly used technology, detailing the most frequently employed variations of whole genome bisulfite sequencing (WGBS) such as classical MethylC-seq and post bisulfite adaptor tagging (PBAT) as well as approaches to focus on parts of the genome through Reduced Representation Bisulfite Sequencing (RRBS). While WGBS comes at a significant higher cost, requires much more extensive computational resources and many sequencing reads will - due to the evolutionary depletion of CpGs in mammals - not contain a single methylation position, it allows to assess all CpGs in a genome. On the other hand, RRBS is much more cost-effective but requires knowledge about the expected genomic context and CpG density to identify the best suited pair of restriction enzymes for selection of these regions. It should be noted that data from human studies points to enhancers as being significantly enriched for inter-individually variable and exposure sensitive DNA methylation patterns rather than CpG island containing promoters, which remain in most cases unmethylated (Gu et al., 2016, Lea et al., 2018). This class of regulatory elements is however difficult to enrich by RRBS.

There are a number of challenges associated with DNA methylation analysis in general and especially in the context of ecology and evolution, which the authors well elaborate on and provide practical solutions. These include notably the required sample size that is required to obtain statistical meaningful results as well as important considerations on key parameters such as for example balancing sequencing depth versus more replicates. One of the most critical parameters as pointed out by the authors, is the type of sample that will be analyzed, a point which cannot be stressed enough. As DNA methylation patterns are cell-type specific, it is important to critically question the relevance of the analyzed sample / tissue type for the scientific question to be addressed in a study. For example, despite its ease of access blood samples might not be an ideal target to analyze behavioral or cognitive traits. A large portion of DNA methylation patterns are similar between different cell-types and tissues, with unmethylated promoters and highly methylated gene bodies and intergenic regions leading to highly significant correlation coefficients between tissues. However, these regions will in most cases also not vary depending on other factors such as environmental conditions, while CpGs displaying variability between tissues are in many cases also those that will be more sensitive to modifications by exposure and correlate with certain phenotypes. Furthermore, if a tissue of relevance can be obtained, its composition will strongly determine the overall DNA methylation patterns. If a specific cell population is of interest, a simple way to avoid confounding is to couple the DNA methylation analysis with cell sorting. Most WGBS protocols are easily compatible with the amount of DNA that can be obtained

2

after cell sorting. If no specific cell population can be preselected, the potential heterogeneity in cellular composition between samples can be identified and corrected for using either cell-type specific reference DNA methylome, or reference-free deconvolution algorithms, which estimate the number of present cell-types (Teschendorff and Zheng, 2017). While reference-based deconvolution algorithms show improved performance in terms of sensitivity and accuracy compared to reference-free methods, reference epigenomes will be rarely available for wild vertebrate species and reference-free methods will nonetheless appropriate to detect large heterogeneity in cell-type composition between the sample and potentially control for them. Of note, one might not always want to control for this heterogeneity as exposure-induced differential cell composition might be of interest itself as we recently demonstrated for in utero exposure to synthetic phenols (Jedynak et al., 2021).

Comprehensive DNA methylation studies using whole genome bisulfite sequencing have been so far scarce in larger sample cohorts mainly due to the cost of sequencing. It should be noted that the number of usable sequences after trimming and mapping for bisulfite converted libraries is substantially lower as for conventional genome sequencing. The use of reference genome sequences from related species due to the absence of a proper reference for the exact species under investigation will further decrease the quantity of uniquely mapped reads. While cost has been the main limiting factor, this year has seen a number of novel high-throughput sequencers being announced and released onto the market that will lead to a significant decrease in costs associated with genome-wide bisulfite sequencing, with DNA methylation analysis already reported for some of them (Lee et al., 2022). Due to the revived competition on the sequencing market, further evolutions in terms of cost and throughput can be expected in the near future enabling the analysis of larger sample sizes. While sequencing protocols based on bisulfite conversion, have proven robust and reproducible with over- or incomplete conversion being a problem in only a minority of cases, alternative protocols making use of enzymatic conversion such as EM-seq (Vaisvila et al., 2021) have been devised and shown similar or better performance as bisulfite-based methods, mainly if very limited amounts of input material were used. Recent developments such as combined genetic and epigenetic sequencing providing simultaneous information on both classic and epigenetically modified nucleotides (Fullgrabe et al., 2023), could simplify DNA methylation analysis in ecology research in the future, especially if combined with long-read sequencing enabling to create a genetic reference genome at the same time as well as the genome-wide map of DNA methylation enabling phasing of genetic and epigenetic variation.

While the analysis of DNA methylation is a first and important step to decipher the mechanisms for altered phenotypes and traits, it will in the future be necessary to analyze the multiple facets of epigenetics to gain a full understanding of the functional impact of epigenetic changes. This includes of course histone modifications as well as non-coding RNAs, but also the oxidative derivatives of cytosine methylation such 5-hydroxymethylation, that is highly prevalent in gene-regulatory elements in some cell-types such as neurons and has an opposite role on gene expression compared to cytosine DNA methylation. Of note, bisulfite treatment is unable to differentiate between DNA methylation and hydroxymethylation and alternative technologies are required (Tost, 2022). Other DNA methylation marks such as adenine methylation and methylation of RNAs (epitranscriptomics) are currently intensively investigated for their gene regulatory potential, but also for their ability to be modified by environmental exposures and their capacity to memorize these exposures (Wu et al., 2023).

The field of ecology and evolution will strongly benefit from these technological advances and new discoveries on the complex gene regulatory landscape defined by epigenetic modifications in humans and model organisms. Although epigenetic analyses might have unique challenges, the guidelines by Laine et al. (Laine et al., 2022) provide very valuable guidance and avoid frustrating and costly mistakes for scientists, who would like to add an epigenetic dimension to their ecology research to fully apprehend the molecular basis of phenotypic variability, rapid environmental adaptation and evolutionary dynamics in which epigenetic modifications play undoubtedly a major role.

References

Fullgrabe, J., Gosal, W. S., Creed, P., *et al.* 2023. Simultaneous sequencing of genetic and epigenetic bases

in DNA. *Nat Biotechnol* .

Gu, J., Stevens, M., Xing, X.*, et al.* 2016. Mapping of Variable DNA Methylation Across Multiple Cell Types Defines a Dynamic Regulatory Landscape of the Human Genome. *G3 (Bethesda),* 6**,** 973-86.

Jedynak, P., Tost, J., Calafat, A. M.*, et al.* 2021. Pregnancy exposure to synthetic phenols and placental DNA methylation - An epigenome-wide association study in male infants from the EDEN cohort.*Environ Pollut,* 290**,** 118024.

Laine, V. N., Sepers, B., Lindner, M.*, et al.* 2022. An ecologist's guide for studying DNA methylation variation in wild vertebrates. *Mol Ecol Resour* .

Lea, A. J., Vockley, C. M., Johnston, R. A.*, et al.* 2018. Genome-wide quantification of the effects of DNA methylation on human gene regulation. *Elife,* 7.

Lee, H., Krieger, G., Clark, T.*, et al.* 2022. Ultra high-throughput whole-genome methylation sequencing reveals trajectories in precancerous polyps to early colorectal adenocarcinoma.*bioRxiv* **,** 2022.05.30.494076.

Teschendorff, A. E. & Zheng, S. C. 2017. Cell-type deconvolution in epigenome-wide association studies: a review and recommendations.*Epigenomics,* 9**,** 757-768.

Tost, J. 2022. Current and Emerging Technologies for the Analysis of the Genome-Wide and Locus-Specific DNA Methylation Patterns. *Adv Exp Med Biol,* 1389**,** 395-469.

Vaisvila, R., Ponnaluri, V. K. C., Sun, Z.*, et al.* 2021. Enzymatic methyl sequencing detects DNA methylation at single-base resolution from picograms of DNA. *Genome Res* .

Wu, H., Eckhardt, C. M. & Baccarelli, A. A. 2023. Molecular mechanisms of environmental exposures and human disease. *Nat Rev Genet,*24**,** 332-344.