HYGEIA: A NEW FRAMEWORK FOR DATA-DRIVEN DISCOVERY OF DNA METHYLATION PATTERNS

Anonymous authors
Paper under double-blind review

Summary. Epigenetic variability is an essential modulator of phenotypic plasticity. To better understand complex epigenetic signals, we introduce Hygeia – a new framework for discovering DNA methylation patterns in whole-genome bisulfite sequencing (WGBS) data. Hygeia utilises a Bayesian statistical model, designed to match empirically observed methylation patterns. The model selects a regime for the methylation propensity at each cytosine-guanine dinucleotide (CpG) site, with regime changes permitted at any position. Thus, conventional means-based methods are replaced by probability-based MEthylaTion changE pOint Regimes (METEORs). Hygeia fits the model to WGBS data to produce METEOR annotation at the CpG level. We applied Hygeia to WGBS EpiATLAS data (N=445) from the International Human Epigenome Consortium (IHEC) to enrich the EpiATLAS resource with METEOR annotation. Hygeia is packaged as a Nextflow pipeline available on GitHub.

Epigenetic modifications are key modulators of DNA and RNA activity. At the DNA level, the most common modification is the addition of a methyl group to the carbon-5 position of CpGs, giving rise to 5-methylcytosine (5mC) which acts as a signaling module in many biological processes (Bird, 2002). Although binary at the single molecule level, CpG methylation patterns become extremely complex at the cellular and tissue levels and highly dynamic at the temporal and spatial levels, where they shape phenotypic plasticity in health and disease (Robertson, 2005). Measurement and analysis of DNA methylation (DNAm) variability has recently been the focus of intense research. While there is a gold standard for generating and preprocessing sequencing-based methylome data, no such standard has yet been defined for the downstream analysis for the learned representations. Since the first whole human methylomes were published in 2008-09, many analysis methods have been developed that have greatly advanced our understanding about the methylome. WGBS data provide counts of DNA molecules in which methylation is observed for a single CpG site. These counts are provided across all CpG sites, for a given depth per site, i.e. over the number of DNA molecules aligned and analysed for each single CpG site.

Most current methods for the analysis of WGBS data are based on the identification of mean differences only, thus a large number of approaches based on such a principle have been suggested to detect differentially methylated positions (DMPs) and differentially methylated regions (DMRs), see e.g. the review in Shafi et al. (2018), with less work aimed at detecting representations beyond the first moment, such as variability (Teschendorff et al., 2016b;a). The popular BSmooth method (Hansen et al., 2012) smoothes the methylation proportions and then tests for group differences using t-tests for each site, but does not allow for a control of Type I error rates when performing multiple-hypotheses testing across sites or regions. Various beta-binomial models have been suggested (Burger et al., 2013; Feng et al., 2014; Park et al., 2014; Sun et al., 2014; Wu et al., 2015), but they tend to allow for only limited spatial dependence, as do most approaches that rely on logistic regressions (Akalin et al., 2012), linear mixed models (Jaffe et al., 2012) or established statistical tests Stockwell et al. (2014). Our approach generalises methods based on hidden Markov models such as Kuan & Chiang (2012); Yu & Sun (2016); Sun & Yu (2016); Shen et al. (2017); Shokoohi et al. (2019); Molaro et al. (2011); Saito et al. (2014) and hidden Semi-Markov models (Du et al., 2014). In particular, it avoids the limitation of the latter that the sojourn time in a particular state is bounded above by some known (and typically relatively small) constant.

In summary, current methods often fail to: (i) allow for flexible methylation patterns that capture variability beyond the mean methylation level; (ii) take into account that methylation of neighboring sites is correlated but also occasionally changes abruptly; (iii) work with a single replicate and missing reads; or (iv) allow for scalable inference on a genome-wide level. Our Hygeia framework addresses these limitations by developing Bayesian change-point models to capture flexible methylation patterns along with the provision of advanced computational algorithms for efficient analysis

of methylome data. Hygeia provides flexible Bayesian change-point models for DNAm and associated inference algorithms that yield a detailed probabilistic description of methylation signatures. The inferred methylation patterns come with uncertainty quantification and can be leveraged for hypothesis-based discoveries with improved power and false discovery rate (FDR) control.

Detection of diverse DNAm patterns. Hygeia replaces current means-based analytics with more powerful probability-based METEORs. The resulting METEOR annotation can be defined by the user and can be tested for any type of differential methylation patterns, enabling the detection of complex DNA methylation dynamics, including spatial and temporal signatures, all within a single framework. Our method can be used to probabilistically segment the methylome into regions of interest, whilst also incorporating domain or expert knowledge in the specification of different METEORs based on known or expected patterns.

Computational efficiency. State-of-the-art Sequential Monte Carlo (SMC) methods enable efficient Bayesian calibration of change-point models in many application domains, including whole genome analysis (Fearnhead & Clifford, 2003; Fearnhead & Liu, 2007; Fearnhead & Vasileiou, 2009; Whiteley et al., 2010; Caron et al., 2012; Yildirim et al., 2013). Within Hygeia, SMC algorithms permit inference on a genome-wide scale because their computational complexity grows only linearly with the number of CpG sites – an attribute regarded as a great challenge for previous Bayesian approaches.

Statistical assessment and differential DNAm patterns. A series of works in multiple testing (Sun & Cai, 2007; Sun & Tony Cai, 2009; Sun & Wei, 2011; Sun et al., 2015) have established optimal testing procedures that maximise the power subject to a constraint on the FDR in case-control scenarios. These procedures are based on the posterior probability of the latent signal, and automatically take into account the spatial dependency of the underlying signal process. Our Bayesian inference strategy provides an effective approximation to the optimal procedure that tightly controls the FDR. In contrast, multiple-testing approaches based on p-values can be overly conservative or challenging to derive under dependence assumptions. In a case-control setting, Hygeia obtains greater statistical efficiency by simultaneously modelling the regimes of the case and control groups. In contrast, some existing methods lose efficiency by fitting independent models to each group.

Open source and user-friendliness. Hygeia has been developed on GitHub with a permissive, open-source license to encourage the creation of an open-source community surrounding it. Hygeia is available in a cloud environment using Nextflow and Seqera Platform, providing a user-friendly web-based solution for launching and monitoring Hygeia analyses at scale.

METEOR annotation of the IHEC EpiATLAS The EpiATLAS is the most comprehensive epigenomic resource. It comprises 19566 datasets consisting of six different histone modifications, RNA-seq and DNA methylation, including the largest single collection of WGBS data. The addition of METEOR annotation to 445 WGBS datasets of this resource will enable researchers to investigate the interplay between DNA methylation and other epigenetic modifications in unprecedented detail.

MEANINGFULNESS STATEMENT

Life is a complex biological process defined by high plasticity at various levels, including development, well-being, and aging, to name just a few. On a molecular level, this plasticity can be measured and quantified over time and space, for instance, through the changing patterns of DNA methylation. DNA methylation is one of many epigenetic modifications that act as regulatory modulators at the intersection of genetics, the environment, and disease. Hygeia is a robust statistical framework that facilitates the analysis of such changing DNA methylation patterns at unprecedented granularity and scale, thus providing novel insights into the plasticity of life.

REFERENCES

Altuna Akalin, Matthias Kormaksson, Sheng Li, Francine E Garrett-Bakelman, Maria E Figueroa, Ari Melnick, and Christopher E Mason. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome biology*, 13(10):1–9, 2012.

Adrian Bird. Dna methylation patterns and epigenetic memory. *Genes & development*, 16(1):6–21, 2002.

Lukas Burger, Dimos Gaidatzis, Dirk Schübeler, and Michael B Stadler. Identification of active regulatory regions from DNA methylation data. *Nucleic Acids Research*, 41(16):e155–e155, 2013.

François Caron, Arnaud Doucet, and Raphael Gottardo. On-line changepoint detection and parameter estimation with application to genomic data. *Statistics and Computing*, 22:579–595, 2012.

Yang Du, Eduard Murani, Siriluck Ponsuksili, and Klaus Wimmers. biomvRhsmm: Genomic segmentation with hidden semi-Markov model. *BioMed Research International*, 2014, 2014.

Paul Fearnhead and Peter Clifford. On-line inference for hidden Markov models via particle filters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(4):887–899, 2003.

Paul Fearnhead and Zhen Liu. On-line inference for multiple changepoint problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):589–605, 2007.

Paul Fearnhead and Despina Vasileiou. Bayesian analysis of isochores. *Journal of the American Statistical Association*, 104(485):132–141, 2009.

Hao Feng, Karen N Conneely, and Hao Wu. A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. *Nucleic Acids Research*, 42 (8):e69–e69, 2014.

Kasper D Hansen, Benjamin Langmead, and Rafael A Irizarry. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biology*, 13:1–10, 2012.

Andrew E Jaffe, Peter Murakami, Hwajin Lee, Jeffrey T Leek, M Daniele Fallin, Andrew P Feinberg, and Rafael A Irizarry. Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *International Journal of Epidemiology*, 41(1):200–209, 2012.

Pei Fen Kuan and Derek Y Chiang. Integrating prior knowledge in multiple testing under dependence with applications to detecting differential DNA methylation. *Biometrics*, 68(3):774–783, 2012.

Antoine Molaro, Emily Hodges, Fang Fang, Qiang Song, W Richard McCombie, Gregory J Hannon, and Andrew D Smith. Sperm methylation profiles reveal features of epigenetic inheritance and evolution in primates. *Cell*, 146(6):1029–1041, 2011.

Yongseok Park, Maria E Figueroa, Laura S Rozek, and Maureen A Sartor. MethylSig: a whole genome DNA methylation analysis pipeline. *Bioinformatics*, 30(17):2414–2422, 2014.

Keith D Robertson. Dna methylation and human disease. *Nature Reviews Genetics*, 6(8):597–610, 2005.

Yutaka Saito, Junko Tsuji, and Toutai Mituyama. Bisulfighter: accurate detection of methylated cytosines and differentially methylated regions. *Nucleic Acids Research*, 42(6):e45–e45, 2014.

Adib Shafi, Cristina Mitrea, Tin Nguyen, and Sorin Draghici. A survey of the approaches for identifying differential methylation using bisulfite sequencing data. *Briefings in Bioinformatics*, 19(5):737–753, 2018.

Linghao Shen, Jun Zhu, Shuo-Yen Robert Li, and Xiaodan Fan. Detect differentially methylated regions using non-homogeneous hidden Markov model for methylation array data. *Bioinformatics*, 33(23):3701–3708, 2017.

Farhad Shokoohi, David A Stephens, Guillaume Bourque, Tomi Pastinen, Celia MT Greenwood, and Aurélie Labbe. A hidden Markov model for identifying differentially methylated sites in bisulfite sequencing data. *Biometrics*, 75(1):210–221, 2019.

Peter A Stockwell, Aniruddha Chatterjee, Euan J Rodger, and Ian M Morison. DMAP: differential methylation analysis package for RRBS and WGBS data. *Bioinformatics*, 30(13):1814–1822, 2014.

Deqiang Sun, Yuanxin Xi, Benjamin Rodriguez, Hyun Jung Park, Pan Tong, Mira Meong, Margaret A Goodell, and Wei Li. MOABS: model based analysis of bisulfite sequencing data. *Genome biology*, 15(2):1–12, 2014.

Shuying Sun and Xiaoqing Yu. HMM-Fisher: identifying differential methylation using a hidden Markov model and Fisher's exact test. *Statistical Applications in Genetics and Molecular Biology*, 15(1):55–67, 2016.

Wenguang Sun and T Tony Cai. Oracle and adaptive compound decision rules for false discovery rate control. *Journal of the American Statistical Association*, 102(479):901–912, 2007.

Wenguang Sun and T Tony Cai. Large-scale multiple testing under dependence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):393–424, 2009.

Wenguang Sun and Zhi Wei. Multiple testing for pattern identification, with applications to microarray time-course experiments. *Journal of the American Statistical Association*, 106(493):73–88, 2011.

Wenguang Sun, Brian J Reich, T Tony Cai, Michele Guindani, and Armin Schwartzman. False discovery control in large-scale spatial multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(1):59–83, 2015.

Andrew E Teschendorff, Yang Gao, Allison Jones, Matthias Ruebner, Matthias W Beckmann, David L Wachter, Peter A Fasching, and Martin Widschwendter. DNA methylation outliers in normal breast tissue identify field defects that are enriched in cancer. *Nature Communications*, 7 (1):10478, 2016a.

Andrew E Teschendorff, Allison Jones, and Martin Widschwendter. Stochastic epigenetic outliers can define field defects in cancer. *BMC Bioinformatics*, 17(1):1–14, 2016b.

Nick Whiteley, Christophe Andrieu, and Arnaud Doucet. Efficient Bayesian inference for switching state-space models using discrete particle Markov chain Monte Carlo methods. *arXiv preprint arXiv:1011.2437*, 2010.

Hao Wu, Tianlei Xu, Hao Feng, Li Chen, Ben Li, Bing Yao, Zhaohui Qin, Peng Jin, and Karen N Conneely. Detection of differentially methylated regions from whole-genome bisulfite sequencing data without replicates. *Nucleic Acids Research*, 43(21):e141–e141, 2015.

Sinan Yildirim, Sumeetpal S Singh, and Arnaud Doucet. An online expectation–maximization algorithm for changepoint models. *Journal of Computational and Graphical Statistics*, 22(4):906–926, 2013.

Xiaoqing Yu and Shuying Sun. HMM-DM: identifying differentially methylated regions using a hidden Markov model. *Statistical Applications in Genetics and Molecular Biology*, 15(1):69–81, 2016.