Title

An Integrated Approach to Modeling Methylation Dynamics

Author

farrell, colin patrick

Publication Date

2021

UNIVERSITY OF CALIFORNIA

Los Angeles

An Integrated Approach to Modeling Methylation Dynamics

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Human Genetics

by

Colin Farrell

2021

ABSTRACT OF THE DISSERTATION

An Integrated Approach to Modeling Methylation Dynamics

by

Colin Farrell
Doctor of Philosophy in Human Genetics
University of California, Los Angeles, 2021
Professor Matteo Pellegrini, Chair

DNA methylation, the addition of a methyl group at the fifth carbon of the pyrimidine ring resulting in 5-methylcytosine (5MC), is influenced by the cellular environment; continued exposure to environmental stimuli will result in detectable DNA methylation changes. These detectable changes have been leveraged to develop DNA methylation based predictive models for age and health. DNA methylation is commonly assayed through the use of bisulfite conversion, where unmethylated cytosine is deaminated to form uracial while 5MC is unchanged, followed by high throughput sequencing. Processing of bisulfite sequencing data is computationally demanding due to the asymmetrical nature of bisulifte sequencing data. Chapter 1 introduces a bisulfite sequencing processing platform, BSBolt, that is a signficant improvement over previous tools in terms of processing time and alignment accuracy. BSBolt and targeted bisulfite sequencing are utilized in chapter 2 to look into epigenetic suppression of transgenic t-cell receptor (TCR) expression in adoptive cell transfer therapy. The chapter 2 study shows that accumulation of DNA methylation over the viral vector promoter used to introduce the TCR sequence is associated with decreased expression despite persistence of the TCR sequence over time. Chapter 3 introduces an evolutionary framework, the epigenetic pacemaker (EPM), for modeling epigenetic aging. The EPM is a departure from the penalized regression based approaches broadly used in the field. The EPM attempts to minimize error across the observed methylation profiles rather than age prediction error. This approach allows the EPM to model nonlinear epigenetic aging across human lifespan as shown in chapter 4. Chapter 5 compares the EPM to penalized regression approaches and shows the EPM is more sensitive for detecting biological signals associated with epigenetic aging.

The dissertation of Colin Farrell is approved.

Jae Hoon Sul

Xia Yang

Aldons J. Lusis

Matteo Pellegrini, Committee Chair

University of California, Los Angeles

2021

This work is dedicated to my mother and father, who nurtured and inspired my love of science.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to express my sincere appreciation and gratitude to my advisor, Matteo Pellgrini. Matteo gave me the academic freedom and support I needed to pursue my research goals. He was always available to discuss my research and help guide me through the challenges I encountered.

I am thankful to the members of my dissertation committee, Jake Lusis, Xia Yang, and Jae-Hoon Sul, who have provided valuable feedback on my dissertation work.

I would like to thank members of the Pellegrini lab who have played a key role in dissertation work and have provided valuable input. With gratitude, thank you to Feiyang Ma, Michael Thompson, Anela Tosevska, Adewale Oyetunde, Morselli Marco, Rubbi Liudmilla, and Bryan Nadal. I also want to thank amy research collaborators, many of whom have contributed to my development as a scientist; particularly, Sag Snir who was intimately involved in a majority of my dissertation work.

I am indebted to John Philips for giving me my first opportunity to pursue research full time. Without his continued support I wouldn't be where I am today. His mentorship and friendship have been instrumental in my development as a scientist.

I have been fortunate to have a strong network of friends during my graduate work at UCLA. I would like to thank Vyncent Jones, for spending his valuable time performing code reviews and talking through machine learning problems. I would also like to thank other members of my graduate cohort James Boocock, Alex Chubick and Arun Durvasula. Their friendship during our shared graduate school experience has been indispensable. I will never forget the long, sometimes impassioned, research discussions over a beer (or two).

A big thank you to my family. My brothers Jonathan, Michael and Sean are great friends and trusted voices who I can always reach out to for advice and guidance. My parents Pat and Patti have always encouraged and supported me.

Finally, I would like to express thanks to my wife Mary for her love and constant support.

She has been my sounding board and confidant throughout my research career. She is always there to lift my spirits and will always be my best friend.

Chapter 2 is a version of the article, "**Farrell, C.**, Thompson, M., Tosevska, A., Oyetunde, A. & Pellegrini, M. BiSulfite Bolt: A BiSulfite Sequencing Analysis Platform. *GigaScience* 2020.10.06.328559 (2020). doi:10.1101/2020.10.06.32855."

Chapter 3 is a version of the article, "Nowicki, T. S., **Farrell, C.**, Morselli, M., Rubbi, L., Campbell, K. M., Macabali, M. H., Berent-Maoz, B., Comin-Anduix, B., Pellegrini, M. & Ribas, A. Epigenetic Suppression of Transgenic T-cell Receptor Expression via Gamma-Retroviral Vector Methylation in Adoptive Cell Transfer Therapy. *Cancer Discov.* (2020). doi:10.1158/2159-8290.CD-20-0300."

Chapter 4 is a version of the article, "**Farrell, C.**, Snir, S. & Pellegrini, M. The Epigenetic Pacemaker: modeling epigenetic states under an evolutionary framework. *Bioinformatics* 36, 4662–4663 (2020)."

Chapter 5 is a version of the article, "Snir, S., **Farrell, C.** & Pellegrini, M. Human epigenetic ageing is logarithmic with time across the entire lifespan. Epigenetics 14, 912–926 (2019)."

Chapter 6 is a version of a manuscript under preparation titled, "Snir, S., **Farrell, C.** & Pellegrini, M. Human epigenetic ageing is logarithmic with time across the entire lifespan. Epigenetics 14, 912–926 (2019)."

VITA

2010 B.S. (Chemistry), University of Utah

2013 M.P.H. (Public Health), University of Utah

PUBLICATIONS

Farrell, C., Hu, H. Pu, K, Lapborisuth, K. Snir, S. & Pellegrini, M. The Epigenetic Pacemaker is a more sensitive tool than penalized regression for identifying factors that impact epigenetic aging. *In Preparation*

Farrell, C., Thompson, M., Tosevska, A., Oyetunde, A. & Pellegrini, M. BiSulfite Bolt: A BiSulfite Sequencing Analysis Platform. *GigaScience* 2020.10.06.328559 (2020). doi:10.1101/2020.10.06.328

Farrell, C., Snir, S. & Pellegrini, M. The Epigenetic Pacemaker: modeling epigenetic states under an evolutionary framework. *Bioinformatics* 36, 4662–4663 (2020).

Nowicki, T. S., **Farrell, C.**, Morselli, M., Rubbi, L., Campbell, K. M., Macabali, M. H., Berent-Maoz, B., Comin-Anduix, B., Pellegrini, M. & Ribas, A. Epigenetic Suppression of Transgenic T-cell Receptor Expression via Gamma-Retroviral Vector Methylation in Adoptive Cell Transfer Therapy. *Cancer Discov.* (2020). doi:10.1158/2159-8290.CD-20-0300

Morselli, M., **Farrell, C.**, Rubbi, L., Fehling, H. L., Henkhaus, R. & Pellegrini, M. Targeted bisulfite sequencing for biomarker discovery. Methods (2020). doi:10.1016/ j.ymeth.2020.07.006

Snir, S., **Farrell, C.** & Pellegrini, M. Human epigenetic ageing is logarithmic with time across the entire lifespan. Epigenetics 14, 912–926 (2019).

Orozco, L. D., **Farrell, C.**, Hale, C., Rubbi, L., Rinaldi, A., Civelek, M., Pan, C., Lam, L., Montoya, D., Edillor, C., Seldin, M., Boehnke, M., Mohlke, K. L., Jacobsen, S., Kuusisto, J., Laakso, M., Lusis, A. J. & Pellegrini, M. Epigenome-wide association in adipose tissue from the METSIM cohort. *Human Molecular Genetics* 27, 1830–1846 (2018).

CHAPTER 1

Introduction

DNA methylation refers to a methyl group being added to a DNA base. Two of four DNA bases can be methylated, cytosine and adenine. Cytosine and adenine can both be methylated by the addition of a methyl group to their amino group, N^4-methylcytosine and N^6-methyladenine respectively. Cytosine can also be methylated in at the fifth carbon of the pyrimidine ring resulting in 5-methylcytosine (5MC). DNA methylation, as used colloquially, refers to 5MC specifically. In mammals, DNA methylation primarily occurs in a CpG (Cytosine-phosphate-Guanine) context, but can also occur in a CH (H=A,T,C) context less frequently[7, 13]. DNA methylation is associated with the topological organization of the cell[16, 21]. Specific cell types carry identifiable methylation patterns and loss of function mutations in cDNA Methylase genes prevents cell differentiation[20]. DNA methylation is also influenced by the cellular environment; continued exposure to stimuli will result in detectable DNA methylation changes[17, 11, 10, 5]. These detectable changes are leveraged to develop DNA methylation based predictive models for age[8, 6] and health[12, 14, 2].

DNA methylation is commonly assayed using bisulfite conversion, where unmethylated cytosine is deaminated to form uracial while 5MC is unchanged[4]. Following, PCR amplification, during which uracil is converted to thymine, the amount of DNA methylation at a genomic location can be quantified by calculating the total number of cytosines (methylated bases) to the total number of bases observed at a given genomic location. Bisulfite converted DNA is primarily assayed using two methods, hybridization arrays and parallel sequencing. Hybridization arrays utilize capture probes to assay DNA methylation at a large number of predetermined locations[1, 19]. The measurements made by hybridization arrays are very precise, with thousands of observations per site. However, while hybridization arrays provide precise measurements at a large number of sites they are costly and assaying sites outside the targeted regions is not feasible. Sequencing based approaches in contrast offer greater flexibility to assay regions of interest at a reduced cost, but generate less precise measurements. Additionally, processing of bisulfite sequencing data is computationally demanding due to the asymmetrical structure of bisulfite sequencing data. Following bisulfite conversion the sense and antisense strands are no longer complementary. A bisulfite converted sequencing read must be mapped to the correct genomic location and origin strand.

Regardless of the assay method, once acquired, DNA methylation measurement can be utilized to fit models of aging[8, 6] and health[12, 14, 2]. The first DNA methylation based predictive models were generated to predict age and are referred to as epigenetic clocks. Epigenetic clocks can accurately predict the age of an individual, and the difference between the predicted and expected epigenetic age has been interpreted as a form of age acceleration. Epigenetic age acceleration has been associated with a number of health related outcomes[3, 9], including mortality[15, 18]. Most work in the area to date, has followed a common workflow where a trait of interest is modeled in DNA methylation data using penalized regression. The goal of penalized regression is to minimize the difference between the observed and predicted value of the modeled trait. This approach can generate highly predictive models, but can minimize informative biological signals.

Chapters 2 - 5 are reformatted versions of published works. Chapter 2 introduces a bisulfite sequencing processing platform, BSBolt, that is a generational improvement over previous tools. Chapter 3 is a study looking into epigenetic suppression of transgenic t-cell receptor expression in adoptive cell transfer therapy. Chapter 4 introduces a novel framework for modeling epigenetic aging, the Epigenetic Pacemaker, (EPM). Chapter 5 investigates non-linear epigenetic aging utilizing the EPM model. Chapter 6 compares the EPM to penalized regression methods that are broadly used in the field to fit epigenetic biomarkers.

Bibliography

[1] Tanya Barrett et al. “NCBI GEO: archive for functional genomics data sets—update”. en. In: *Nucleic Acids Res.* 41.D1 (Nov. 2012), pp. D991–D995.

[2] Daniel W Belsky et al. “Quantification of the pace of biological aging in humans through a blood test, the DunedinPoAm DNA methylation algorithm”. en. In: *Elife* 9 (May 2020).

[3] Pierre-Antoine Dugué et al. “DNA methylation-based biological aging and cancer risk and survival: Pooled analysis of seven prospective studies”. en. In: *Int. J. Cancer* 142.8 (Apr. 2018), pp. 1611–1619.

[4] M Frommer et al. “A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands”. en. In: *Proc. Natl. Acad. Sci. U. S. A.* 89.5 (Mar. 1992), pp. 1827–1831.

[5] Eilis Hannon et al. “DNA methylation meta-analysis reveals cellular alterations in psychosis and markers of treatment-resistant schizophrenia”. en. In: *Elife* 10 (Feb. 2021).

[6] Gregory Hannum et al. “Genome-wide methylation profiles reveal quantitative views of human aging rates”. en. In: *Mol. Cell* 49.2 (Jan. 2013), pp. 359–367.

[7] Yupeng He and Joseph R Ecker. “Non-CG Methylation in the Human Genome”. en. In: *Annu. Rev. Genomics Hum. Genet.* 16 (June 2015), pp. 55–77.

[8] Steve Horvath. *DNA methylation age of human tissues and cell types.* 2013.

[9] Rae-Chi Huang et al. “Epigenetic Age Acceleration in Adolescence Associates With BMI, Inflammation, and Risk Score for Middle Age Cardiovascular Disease”. en. In: *J. Clin. Endocrinol. Metab.* 104.7 (July 2019), pp. 3012–3024.

[10] Juliana Imgenberg-Kreuz et al. *P89 Epigenome-wide association study reveals differential DNA methylation in systemic lupus erythematosus patients with a history of ischemic heart disease*. 2020.

[11] Leanne K Küpers et al. "Meta-analysis of epigenome-wide association studies in neonates reveals widespread differential DNA methylation associated with birthweight". en. In: *Nat. Commun.* 10.1 (Apr. 2019), p. 1893.

[12] Morgan E Levine et al. "An epigenetic biomarker of aging for lifespan and healthspan". en. In: *Aging* 10.4 (Apr. 2018), pp. 573–591.

[13] Ryan Lister et al. "Human DNA methylomes at base resolution show widespread epigenomic differences". en. In: *Nature* 462.7271 (Nov. 2009), pp. 315–322.

[14] Ake T Lu et al. "DNA methylation GrimAge strongly predicts lifespan and healthspan". en. In: *Aging* 11.2 (Jan. 2019), pp. 303–327.

[15] Riccardo E Marioni et al. "DNA methylation age of blood predicts all-cause mortality in later life". en. In: *Genome Biol.* 16 (Jan. 2015), p. 25.

[16] Stephan Nothjunge et al. *DNA methylation signatures follow preformed chromatin compartments in cardiac myocytes*. 2017.

[17] Luz D Orozco et al. "Epigenome-wide association in adipose tissue from the METSIM cohort". en. In: *Hum. Mol. Genet.* 27.14 (July 2018), p. 2586.

[18] Laura Perna et al. *Epigenetic age acceleration predicts cancer, cardiovascular, and all-cause mortality in a German case cohort*. 2016.

[19] Ruth Pidsley et al. "Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling". en. In: *Genome Biol.* 17.1 (Oct. 2016), p. 208.

[20] Zachary D Smith and Alexander Meissner. "DNA methylation: roles in mammalian development". en. In: *Nat. Rev. Genet.* 14.3 (Mar. 2013), pp. 204–220.

[21] Elena K Stamenova et al. “The Hi-Culfite assay reveals relationships between chromatin contacts and DNA methylation state”. en. Nov. 2018.

CHAPTER 2

BiSulfite Bolt: A BiSulfite Sequencing Analysis Platform

Colin Farrell[1], Michael Thompson[2], Anela Tosevska[2], Adewale Oyetunde[2], Matteo Pellegrini[2,3]

[1]Department of Human Genetics, University of California, Los Angeles, CA, USA;

[2]Dept. of Molecular, Cell and Developmental Biology; University of California, Los Angeles, CA 90095, USA;

[3]Corresponding Author, matteop@mcdb.ucla.edu

Background: Bisulfite sequencing is commonly employed to measure DNA methylation. Processing bisulfite sequencing data is often challenging due to the computational demands of mapping a low complexity, asymmetrical library and the lack of a unified processing toolset to produce an analysis ready methylation matrix from read alignments. To address these shortcomings, we have developed BiSulfite Bolt (BSBolt); a fast and scalable bisulfite sequencing analysis platform. BSBolt performs a pre-alignment sequencing read assessment step to improve efficiency when handling asymmetrical bisulfite sequencing libraries.

Findings: We evaluated BSBolt against simulated and real bisulfite sequencing libraries. We found that BSBolt provides accurate and fast bisulfite sequencing alignments and methylation calls. We also compared BSBolt to several existing bisulfite alignment tools and found BSBolt outperforms Bismark, BSSeeker2, BISCUIT, and BWA-Meth based on alignment accuracy and methylation calling accuracy.

Conclusion: BSBolt offers streamlined processing of bisulfite sequencing data through an integrated toolset that offers support for simulation, alignment, methylation calling, and data aggregation. BSBolt is implemented as a python package and command line utility for flexibility when building informatics pipelines. BSBolt is available at https://github.com/NuttyLogic/BSBolt under a MIT license.

2.1 Findings

Background

DNA methylation, the epigenetic modification of cytosine by the addition of a methyl group to the fifth carbon of the cyclic backbone, is a widely studied epigenetic mark associated with gene regulation[24, 25] and numerous biological processes [7, 18, 23]. High throughput sequencing combined with bisulfite conversion is a broadly used method for profiling DNA methylation genome wide[16][17]. Treatment of DNA with sodium bisulfite results in unmethylated cytosines being deaminated to uracil, and converted to thymine through PCR amplification, while methylated cytosine, guanine, thymine, and adenine remain unchanged [4]. The methylation status of an individual site or region can be assessed by looking at the number bisulfite converted bases relative to the total number of observed bases. Amongst eukaryotic organisms the majority of genomic cytosines are unmethylated [4, 3, 15]. As a consequence, bisulfite sequencing reads originating from the same location but opposite strands are generally no longer complementary. Additionally, when the PCR product of the original bisulfite converted sequence is considered, sequencing reads can be aligned in different orientations within the same strand. Given the asymmetrical nature of bisulfite sequencing libraries and the large number of potential mismatches between the read sequence and the reference the use of a traditional alignment tool would produce low quality alignments.

Bisulfite sequencing alignment tools Bismark[9], BS-Seeker2[9, 6], and BWA-Meth[19] successfully adopted a three-base alignment strategy wrapped around established read aligners such as Bowtie2[11, 10] and BWA-MEM[11], to accurately align bisulfite sequencing reads. In this strategy, an alignment index or multiple alignment indices are generated against each bisulfite converted reference strand. Relative to the reference, the bisulfite sense strand is the reference with all cytosines converted to thymine and the antisense strand is the reference sequence with all guanines converted to adenine. Before alignment, input reads are in silico bisulfite converted so any methylated or incompletely converted bases are converted to remove mismatches relative to the bisulfite reference. Reads are then aligned using the wrapped read alignment tool and the output alignments are integrated together with the original read sequence to form a consensus alignment file. During the generation of a consensus alignment file BS-Seeker2 and Bismark call contextual methylation, where CG methylation is reported distinctly from CH (H=A,C,T) methylation, for every aligned base within an alignment. The regional methylation information provided within alignment calls can provide important context about the epigenetic organization of a genome and the reorganization that occurs in response to disease [8, 5, 14]. Methylation calls from aligned reads can also be leveraged to assess the bisulfite conversion status of a read. A high proportion of observed methylated CH sites relative to the total number of observed CH indicates a read that was incompletely bisulfite converted as the majority of CH sites are expected to be unmethylated.

The three base alignment strategy as implemented by BSSeeker2 and Bismark has several limitations. Both tools carry out multiple intermediary alignments to separate alignment indices representing different reference conversion patterns and then integrate intermediate alignments together into a consensus alignment file. Reads with multiple alignments within an intermediate alignment file or across multiple intermediate alignment files are discarded; only reads that align uniquely within a single intermediate alignment are reported. In an effort to reduce the number of reads that align across alignment indices both BSSeeker2 and

Bismark have strict default alignment parameters. In addition to being computationally demanding, this implementation can also reduce the number of valid alignments reported, as only the highest quality, unique alignments are output. BWA-Meth resolves this issue by performing alignment to a single bisulfite converted alignment index and processing reads on the fly; but, does not return the read level methylation calls or bisulfite conversion assessment provided by Bismark and BSSeeker2. Additionally, when performing bisulfite sequencing alignment the read conversion pattern is dependent on whether the sequenced DNA fragment is representative of the original DNA sequence or its PCR product. In a directional bisulfite sequencing library only DNA representative of the original DNA fragment is sequenced so the bisulfite conversion pattern is known. In an undirectional library, DNA representative of the original DNA fragment and its PCR product is sequenced so a cytosine to thymine or a guanine to adenine conversion is possible. BS-Seeker2 and Bismark handle undirecitonal libraries by converting input reads using both conversion patterns. This approach doubles the number of reads that must be aligned and generates input reads that will not be represented in the alignment index. BWA-Meth does not support alignment of undirectional libraries.

Here we present BiSulfiteBolt (BSBolt), a bisulfite sequencing platform designed to be fast and scalable while als providing the same read-level methylation calls and quality metrics of BS-Seeker2 and Bismark to preserve compatibility with existing analysis tools. BSBolt alignment is built on a forked version of BWA-MEM[13, 11] and HTSLIB[13] with bisulfite specific sequencing logic integrated directly into the alignment process. BSBolt incorporates a pre-alignment read assessment step to assess the correct conversion pattern when aligning undirectional libraries. This eliminates the needs to perform multiple alignments for the same read, improving performance. Additionally, as the output alignment structure is slightly different between each bisulfite alignment wrapper, each tool implements its own methylation calling utility and output format. BSBolt includes a rapid and multi-threaded methylation caller, that outputs methylation calls in CGmap or bedGraph format implemented by BSSeeker2 and Bismark respectively. We show that BSBolt alignments and methylation calling is considerably faster and more accurate than these other bisulfite sequencing alignment wrappers. Additionally, we compare BSBolt to another high performance bisulfite sequencing platform BISCUIT[2]. BISCUIT also incorporates bisulfite specific alignment logic directly into the alignment process, but doesn't support read level methylation calling or bisulfite conversion assessment during alignment. Despite this, we show that BSBolt offers comparable, or faster, performance. Additionally, to facilitate end to end processing of bisulfite sequencing data BSBolt includes utilities for read simulation utility and aggregation of methylation call files into a consensus matrix.

2.2 Methods

BSBolt Workflow

BSBolt Alignment

BSBolt incorporates bisulfite alignment logic directly within a forked version of BWA-MEM. BSBolt is designed around a single Burrows-Wheeler Transform (BWT) FM-index constructed from both bisulfite converted reference strands. BSBolt utilizes a three base alignment strategy where input reads sequences are fully in silico converted before alignment. In this case of undirectional libraries, where a cytosine to thymine or guanine to adenine conversion if possible, BSBolt first analyzes the read base composition. A read, or read pair, with a low proportion of observed cytosines compared to guanine (0.1 by default) will be preferentially aligned with a cytosine to thymine conversion pattern and vice versa. If it is unclear what conversion pattern should be used, both conversion patterns are aligned and the conversion pattern with the highest total alignment score is output. The converted read sequence is aligned using BWA-MEM to the bisulfite FM-index. The resulting alignments are then modified so reads mapping to the sense reference strand are reported as sense reads and the anti-sense reference reported as antisense reads regardless of mapping orientation.

The mapping quality of an alignment is assessed by mapping uniqueness using standard BWA-MEM scoring criteria. Additionally, an alignment with alternative alignments on a different bisulfite reference strand is further penalized for being bisulfite ambiguous. Read variation and methylation calls are then made for alignments meeting scoring thresholds using the original read sequence and an unconverted reference sequence. If a difference between the alignment and reference is explainable by bisulfite conversion a methylation call is made for the aligned base; otherwise, reference variation is reported. When calling methylation values, the context of the methylatable base is considered by capturing the local reference context (ie CG or CH). The methylation calls are output as a Sequence Alignment/Map (SAM) flag mirroring the BWA-MEM MD flag. Typically, the majority of CH sites are unmethylated so the expectation is that the majority of CH sites within a read, or read pair, are bisulfite converted. After calling read level methylation this information is leveraged to assess the bisulfite conversion status of the read across all aligned bases within the read, or read pair. The conversion status of the read is conveyed as a SAM flag in the output alignment. Output alignments are then compressed and written to a bam file natively.

BSBolt Methylation Calling

BSBolt includes an optimized methylation calling utility that takes advantage of the BSBolt alignment file structure to rapidly call site methylation. The calling procedure proceeds as follows. A read pileup is created using samtools[13], and initialized using pysam[21], for each reference contig with aligned reads. Methylation calls are made for all methylatable bases, or only CG sites, using all reads that pass user specified quality metrics. Methylation values for reference guanine nucleotides are made for reads aligned to the antisense strand and calls for reference cytosine nucleotides are made for reads aligned to the sense strand. This call strategy decreases methylation calling time, as information about the origin strand can be quickly interpreted. Methylation calls are then output in the CGmap file format implemented by BSSeeker2. To aggregate several call files together into a consensus matrix BSBolt includes a rapid and efficient matrix aggregation utility. Bisulfite sequencing techniques often capture methylation sites unevenly, so making a combined matrix of all sites observed across every call file can be inefficient and produce large sparse matrices. BSBolt utilizes an iterative matrix assembly method where individual CGmap files are iterated through to count how often individual sites appear at or above a user specified coverage threshold. If a site is observed in a set proportion of the CGmap files the site is included in the consensus matrix. This process is parallelizable across several threads for efficiency. BSBolt supports output of matrices containing methylation values and counts of methylated and total bases at each site.

BSBolt Simulation

BSBolt Simulate utilizes a modified version of WGSIM[12] wrapped with python to simulate bisulfite converted reads with site specific methylation information incorporated across reads. Given a reference sequence global methylation values are set by randomly selecting a methylation value for all methylatable bases depending on context (CG or CH) or by passing a methylation profile in the form of a CGmap file. Reads are then simulated by randomly selecting a genomic position within a reference sequence, sampling the reference sequence at set read length, and insert size for paired end reads, then incorporating sequencing error and genetic variation. The origin strand, and conversion pattern if simulating undirectional reads, is then randomly selected. At every methylatable base within a read the methylation status of the base is set by the probability of observing a methylated base given the reference methylation value. The mapping location, methylation status, and origin bisulfite strand are attached as a fastq comment and output along with the bisulfite converted read sequence and base call qualities. The number of methylated and unmethylated bases covering each methylation site are output as a serialized python object at the end of the simulation.

Tool Comparisons

BSBolt (v1.4.4), BISCUIT (v0.3.16.20200420), BSSeeker2 (v2.1.8), BWA-Meth (v0.2.2), and Bismark (v0.22.3) were used for comparisons with both real and simulated bisulfite sequencing data. All comparisons were performed on a compute node with XEON X5650 six core (twelve thread) processor (48GB ram) running centos (v6.10). Each tool was provided with 12 compute threads if supported. Default alignment parameters were used unless library specific alignment options were necessary to support the simulated library type. Uncompressed alignment outputs were compressed using samtools (v1.9) before being written to disk. Samtools and BSBolt were provided with two compression threads to minimize any alignment bottlenecks (S. Figure 1). If supported, methylation calls were only made using reads with a mapping quality higher than 20.

Simulated Bisulfite Library Comparisons

A simulation reference genome was created by sampling approximately 2Mb from each chromosome in the human reference genome (hg38) excluding alternative and sex chromosomes. Briefly, 50bp tiles were randomly sampled from a reference chromosome and included in the simulation reference if the tile contained less than 10 ambiguous bases. The first 10kb of the simulated chr1 was duplicated and added as an additional contig. A series of directional and undirectional bisulfite sequencing libraries were then simulated using BSBolt at various read lengths, read depths, and read qualities with random methylation profiles (Table 1). Alignment and methylation calling tools for each package were compared by aligning a simulation library, sorting the alignment file if necessary, and calling methylation values. Each simulation library was processed by each comparison package sequentially in random order on the same compute node. Read alignments were evaluated by the alignment location and strand. An on-target alignment was defined as a read where 95% of the aligned bases were mapped within the simulated region and mapped to the correct origin strand. An alignment was considered off-target if fewer than 5% of the aligned bases were mapped to the simulation region, the aligned strand of origin was incorrect or flagged as a quality control failure. Accuracy of the CpG methylation calls were evaluated by comparing the called methylation value with the simulated value.

Targeted Bisulfite Library Comparisons

We next utilized publicly available targeted bisulfite sequencing data (GSE152923) generated from peripheral blood mononuclear cells of four individuals [Citation error].The libraries were generated using the SureSlectXT Methyl-Seq (Aligent) kit and three sequencing libraries were generated for each individual with varying levels of input DNA (1000ng, 300-1000ng, and 150ng-300ng). Each library was sequenced (100bp, paired end) on an Illumina NovaSeq generating an average of 144.1 million (118.5 - 230.5) paired end reads. In addition to the sequencing data, methylation measurements were generated using the Infinium MethylationEPIC array (Illumina) for all four individuals. Whole genome bisulfite alignment indices were generated using hg38 for each bisulfite sequencing package. Every sequencing library was aligned and processed using the same workflow. Alignment files were generated, duplicate reads were marked using samtools (v1.9), and methylation values were called. Each alignment and methylation calling workflow was given a maximum runtime of 288 hours. If an alignment was incomplete at the end of 288 hours, duplicate read marking and methylation calling was performed on the reads aligned during the 288 hour limit. Methylation calls made for CpG sites with more than five reads covering a site were then compared with array methylation values from the same biological sample.

2.3 Results

BSBolt was the fastest alignment tool across all simulation conditions, aligning close to 2.29 million reads per minute on average (Figure 2A). BSBolt was approximately 40% faster than the next fastest alignment tool, BISCUIT. When looking at alignment performance by library type, BISCUIT exhibited similar performance to BSBolt when aligning directional reads, but was approximately 229% slower aligning undirectional libraries (Figure 2A). BSSeeker2, BWA-Meth, and Bismark were slower than both BSBolt and BISCUIT when aligning all library types (Figure 2A). BSBolt and BISCUIT aligned the majority of simulated reads across all conditions (>99%) with high accuracy (>99%). BWA-Meth aligned the majority of reads accurately for directional libraries, but as undirectional libraries are unsupported, BWA-Meth undirectional alignments had low mappability ($\mu = 0.724$) and a low proportion of aligned reads were on target ($\mu = 0.706$). BSSeeker2 and Bismark exhibited the lowest average mappability across all simulation conditions at 93.6% and 86.9% respectively but the output alignments were generally accurate (Figure 2B). Moreover, BSSeeker2 and Bismark aligned a low percentage of the simulated reads, 65.3% and 42.4% respectively, when the simulated sequencing error and genetic variation was increased from 0.05% to 2% (S. Table 1). Bismark and BSSeeker2 both discard base call quality information when aligning reads so the low mappability with error prone reads is expected.

BSBolt methylation calling was significantly faster than all other tools, with a roughly 11 fold performance advantage over the next fastest tools, BISCUIT and BWA-Meth. BSeeker2 and Bismark were considerably slower and exhibited a strong relationship between call time and the number of simulated reads (Figure 2C). We also looked at the mean absolute error (MAE) between the number of reads simulated at a given position and the number of reads utilized by each tool to call methylation. BSBolt had the lowest average MAE (0.11 reads) followed by BISCUIT (0.70 reads) and Bismark (0.76 reads). BWA-Meth and BSSeeker2 exhibited high coverage MAE at 6.12 and 8.69 reads respectively. While the BSSeeker2 coverage MAE was high it was not strand biased and the methylation level MAE was small, 0.024. By contrast, the methylation calls made by BWA-Meth were strand biased as shown by the methylation value MAE, 0.255. Overall, BSBolt had the lowest observed methylation level MAE (0.002) followed by BISCUIT (0.013) and Bismark (0.024) (Figure 2D).

The performance of each tool with the targeted bisulfite sequencing libraries largely mirrored the results with the simulation data. However, even though the targeted libraries are directional, BSBolt outperformed BISCUIT aligning an average of 663k reads per minute compared with 637k (Figure 3A). BSSeeker2 failed to align three sequencing libraries within the 288 hour alignment limit, aligning only 78% of reads on average. BSBolt was the fastest methylation calling tool, calling CpG methylation in just 4.35 minutes on average (Figure 3B). We then compared the absolute differences between the sequencing and Illumina EPIC array calls made for the same biological sample, excluding BSSeeker2 alignments as three alignments were incomplete. The absolute differences for all comparisons were combined by tool and binned by effective read coverage, or the number of reads used to call the methylation value (Figure 3C). The called methylation values were highly correlated with the sites called on the EPIC array across all alignment tools (Pearson's r=.92-98, S. Table 2), as previously reported [22]. Unsurprisingly, as sequencing depth increases the observed mean absolute deviation decreases for all tools. At sequencing depths above 40 reads per CpG BSBolt has the smallest absolute deviation between the sequencing and array calls. Note, due the design of the targeted bisulfite libraries, DNA from one origin strand is preferentially captured over a given region. As a result, the strand bias of the BWA-Meth methylation caller didn't noticeably impact the methylation calls.

2.4 Discussion

Both BSBolt and BISCUIT are significantly faster at bisulfite read alignment while also being more accurate on average than BSSeeker2, Bismark, and BWA-Meth. BSBolt offered marginal performance improvement over BISCUIT with real directional bisulfite libraries, but a large performance gain for the simulated undirectional libraries due to the implementation of a pre-alignment sequencing assessment step. In addition to aligning each read, BSBolt calls contextual read level methylation and assesses read bisulfite conversion, generating alignment information similar to Bismark and BSSeeker2. Importantly, as Bismark and BSSekeer2 have been widely adopted by the community at large it is important to provide the same alignment information to preserve compatibility with downstream tools. BISCUIT offers support for read bisulfite conversion assessment but it is implemented as post-alignment utility.The BSBolt methylation caller was significantly faster than other tools while also providing more accurate methylation calls. Much of this improvement can be attributed to the structuring read alignment before output; by modifying the alignment strand to reflect the bisulfite origin strand methylation calls can be made rapidly without the need to perform additional formatting.

BSBolt is implemented as a python package installable through the python package index[20] and the Anaconda package manager[1]. In addition to a fully command line interface each BSBolt module can be executed natively as an object in a python (>3.6) environment; providing flexibility for informatics pipelines. BSBolt is available at https://pypi.org/project/BSBolt/ and is released under the MIT license.

2.5 Availability and requirements

Project Name: BSBolt

Project Home Page: https://github.com/NuttyLogic/BSBolt

Operating system(s): Platform Independent

Programming language: Python $\geq$ 3.6

Other requirements: numpy$\geq$1.16.3, tqdm$\geq$4.31.1

License: MIT

RRID: SCR019080

Acknowledgments and Funding

This work was supported by theNational Institutes of Health (T32CA201160 to C.F.).

Supplementary Information

Analysis Code: https://github.com/NuttyLogic/BSBoltManuscript

Supplemental Table 1: Simulated Bisulfite Sequencing Library Run Stats

https://github.com/NuttyLogic/BSBoltManuscript/blob/

master/BSBolt%20Supplemental%20Table%201.xlsx

Supplemental Table 2: Targeted Bisulfite Alignment Stats

https://github.com/NuttyLogic/BSBoltManuscript/blob/

master/BSBolt%20Supplemental%20Table%202.xlsx

Supplemental Figure1: Samtools BAM Conversion Thread Comparisons

Supplemental Figure2: BSBolt Performance Characteristics on 150bp Simulated Libraries

Data Availability

Targeted bisulfite sequencing and EPIC array data deposited in GEO, GSE152923. The pipeline used to simulate bisulfite sequencing libraries is deposited in the analysis repository.

Author Contributions

CF developed BSBolt, performed manuscript data analysis, and drafted manuscript. CF, AO, AT, and MT evaluated and tested BSBolt. CF, AO, AT, MT and MP edited manuscript. MP and CF conceptualized the project.

This work used computational and storage services associated with the Hoff-

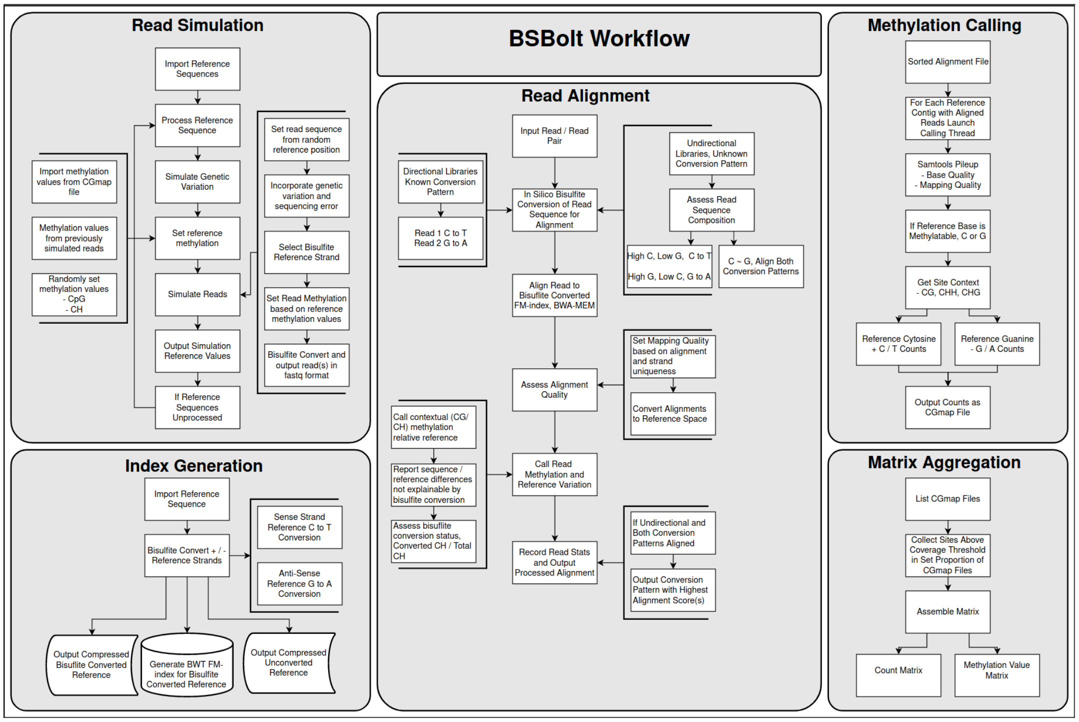

BSBolt Workflow
Read Simulation
Import Reference Sequences
Process Reference Sequence
Simulate Genetic Variation
Set reference methylation
Simulate Reads
Output Simulation Reference Values
If Reference Sequences Unprocessed
Import methylation values from CGmap file
Methylation values from previously simulated reads
Randomly set methylation values - CpG - CH
Set read sequence from random reference position
Incorporate genetic variation and sequencing error
Select Bisulfite Reference Strand
Set Read Methylation based on reference methylation values
Bisulfite Convert and output read(s) in fastq format
Index Generation
Import Reference Sequence
Bisulfite Convert + / - Reference Strands
Sense Strand Reference C to T Conversion
Anti-Sense Reference G to A Conversion
Output Compressed Bisuflite Converted Reference
Generate BWT FM-index for Bisulfite Converted Reference
Output Compressed Unconverted Reference
Read Alignment
Input Read / Read Pair
Directional Libraries Known Conversion Pattern
Read 1 C to T Read 2 G to A
In Silico Bisulfite Conversion of Read Sequence for Alignment
Undirectional Libraries, Unknown Conversion Pattern
Assess Read Sequence Composition
High C, Low G, C to T
High G, Low C, G to A
C ~ G, Align Both Conversion Patterns
Align Read to Bisuflite Converted FM-index, BWA-MEM
Set Mapping Quality based on alignment and strand uniqueness
Convert Alignments to Reference Space
Assess Alignment Quality
Call contextual (CG/CH) methylation relative reference
Report sequence / reference differences not explainable by bisulfite conversion
Assess bisuflite conversion status, Converted CH / Total CH
Call Read Methylation and Reference Variation
If Undirectional and Both Conversion Patterns Aligned
Output Conversion Pattern with Highest Alignment Score(s)
Record Read Stats and Output Processed Alignment
Methylation Calling
Sorted Alignment File
For Each Reference Contig with Aligned Reads Launch Calling Thread
Samtools Pileup - Base Quality - Mapping Quality
If Reference Base is Methylatable, C or G
Get Site Context - CG, CHH, CHG
Reference Cytosine + C / T Counts
Reference Guanine - G / A Counts
Output Counts as CGmap File
Matrix Aggregation
List CGmap Files
Collect Sites Above Coverage Threshold in Set Proportion of CGmap Files
Assemble Matrix
Count Matrix
Methylation Value Matrix

Figure 1: BSBolt Workflows

BSBolt is implemented as a series of discrete modules for read simulation, index generation, read alignment, methylation calling, and matrix aggregation. All BSBolt modules can be run using a command line interface or within a python (> 3.6) environment natively.

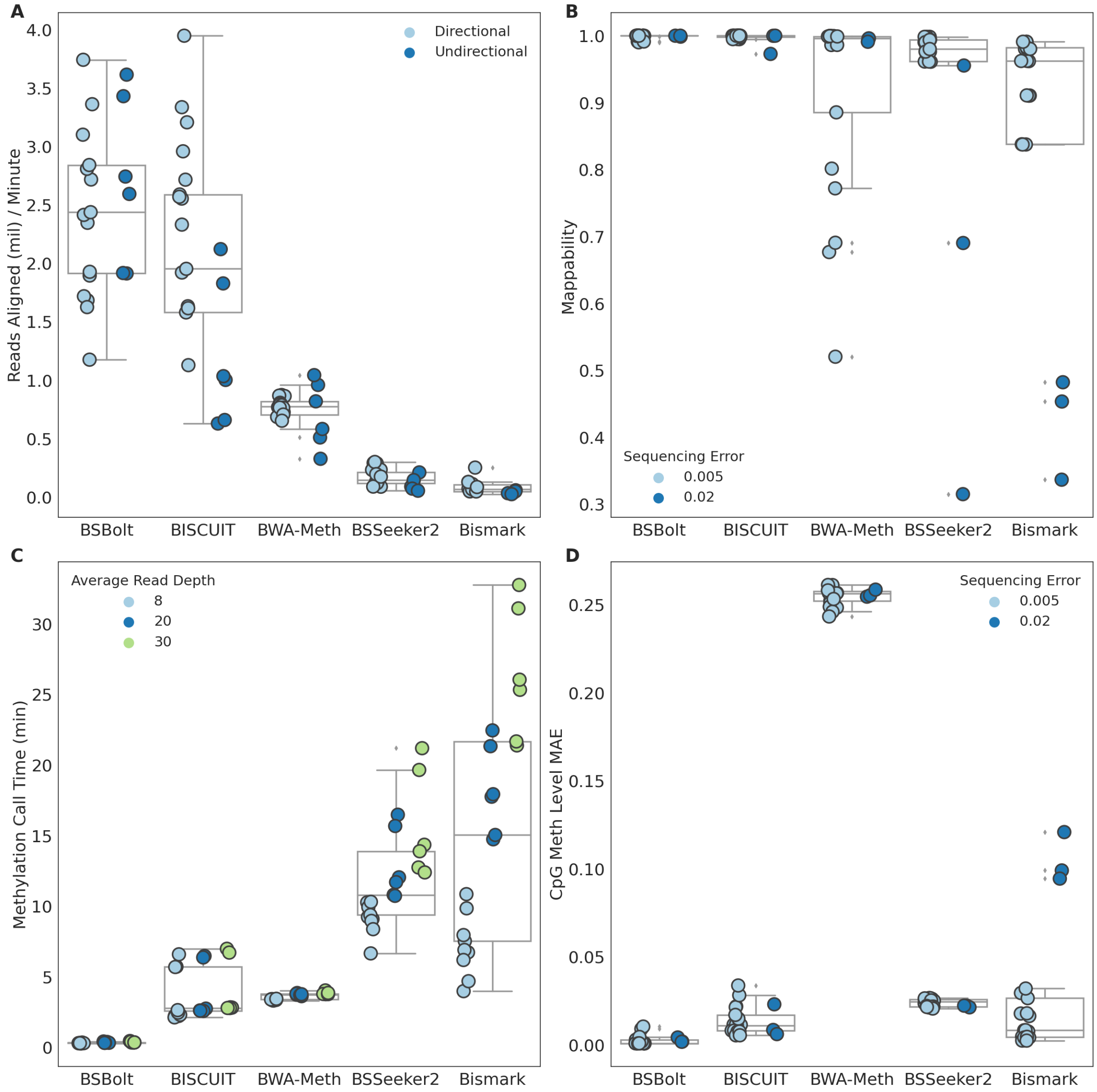
man2 Shared Cluster provided by UCLA Institute for Digital Research and Education's Research Technology Group.

*

Table 1: Simulated Bisulfite Sequencing Library Parameters:

The parameters used to simulate libraries using BSBolt for tool comparisons. All simulations were carried out at read lengths of 50, 100, and 150 base pairs.

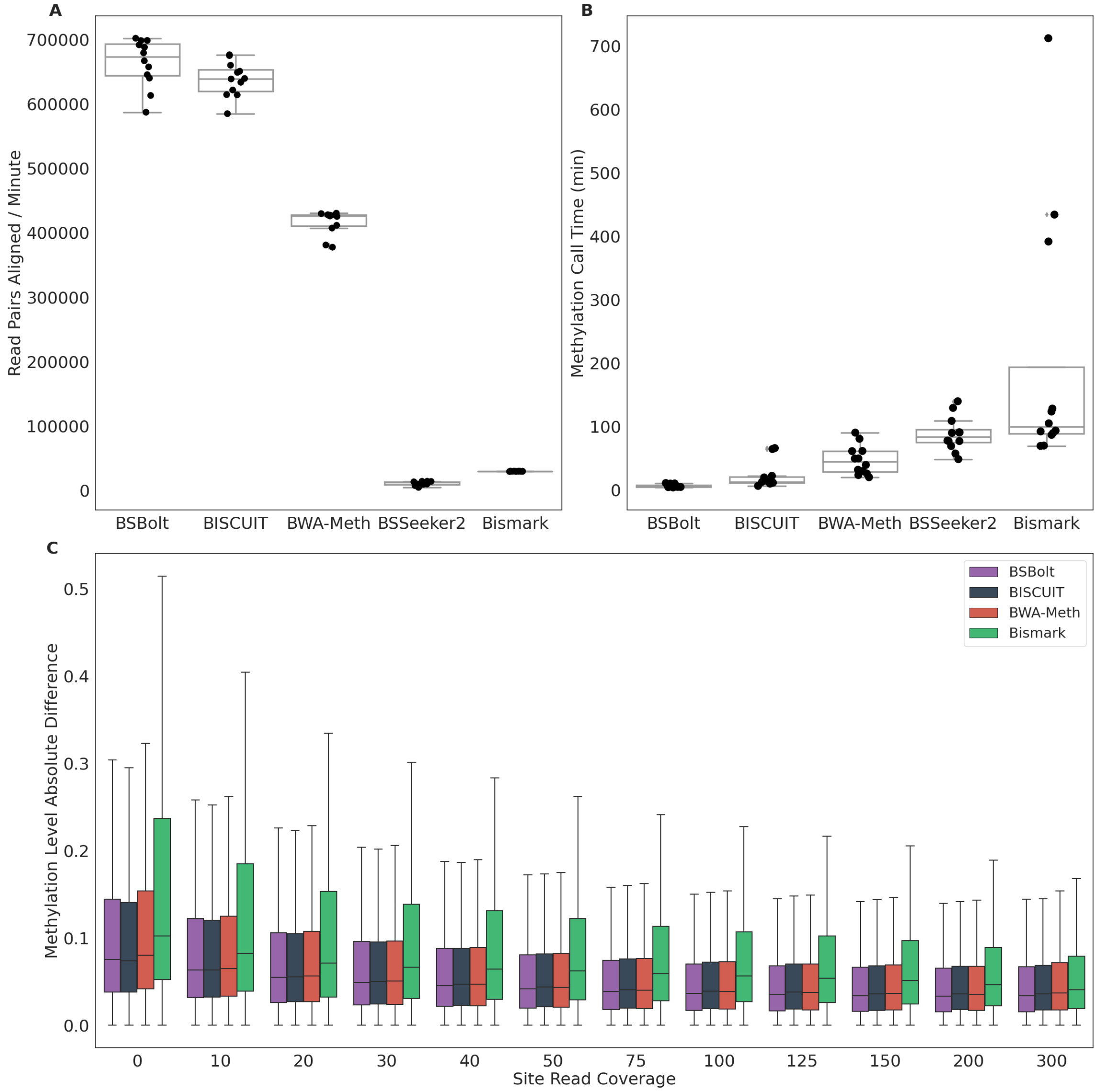
Average Read Depth	**Mutation Rate**	**Sequencing Error**	**Sequencing Type**	**Library Type**
30	0.005	0.005	Paired End	Undirectional
30	0.005	0.005	Single End	Undirectional
20	0.005	0.005	Paired End	Directional
20	0.005	0.005	Single End	Directional
8	0.005	0.005	Paired End	Directional
8	0.005	0.005	Single End	Directional
8	0.01	0.02	Paired End	Directional



*

Figure 2: Simulated Bisulfite Sequencing Library Performance

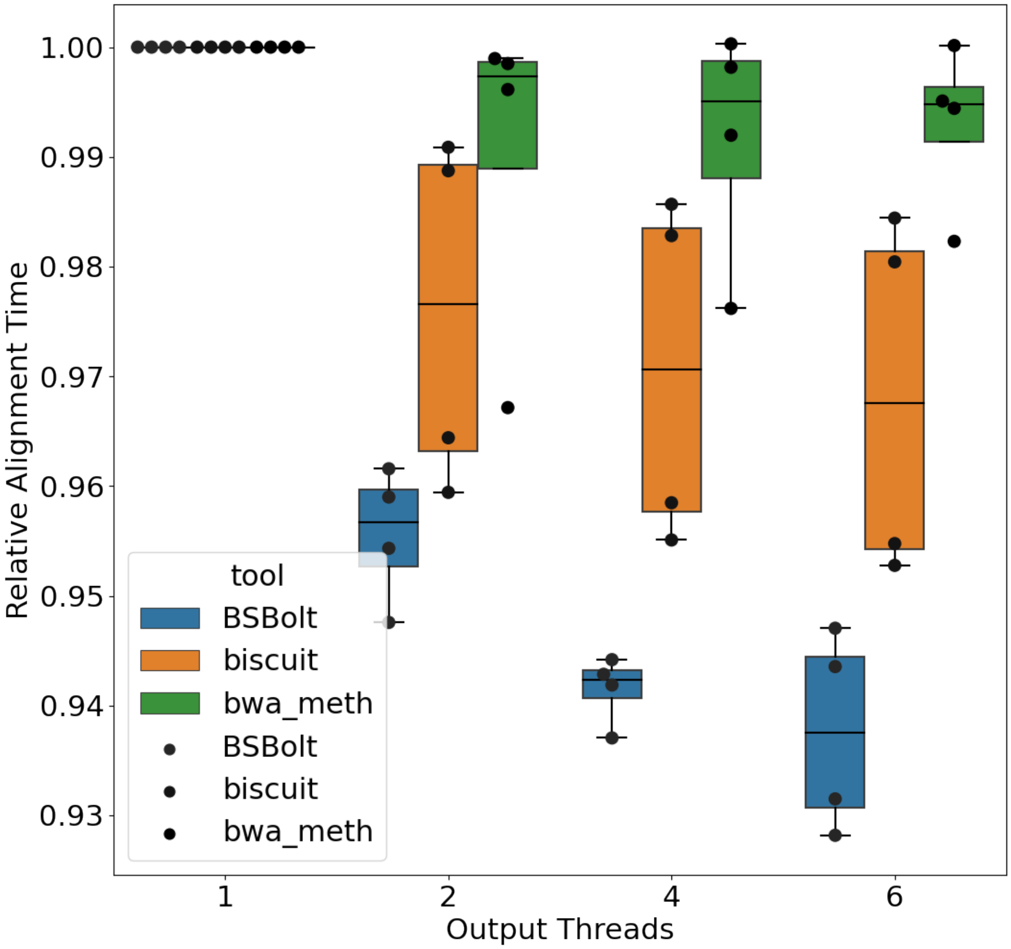
(A) Reads aligned per minute for each bisulfite alignment tool. (B) Proportion of simulated reads mapped during alignments. Note, BWA-Meth does not support undirectional library alignment resulting in low mappability for undirectional libraries. (C) Methylation call time (min) for each alignment tool. (D) Mean Absolute Error (MAE) observed between the simulated and called methylation value.



*

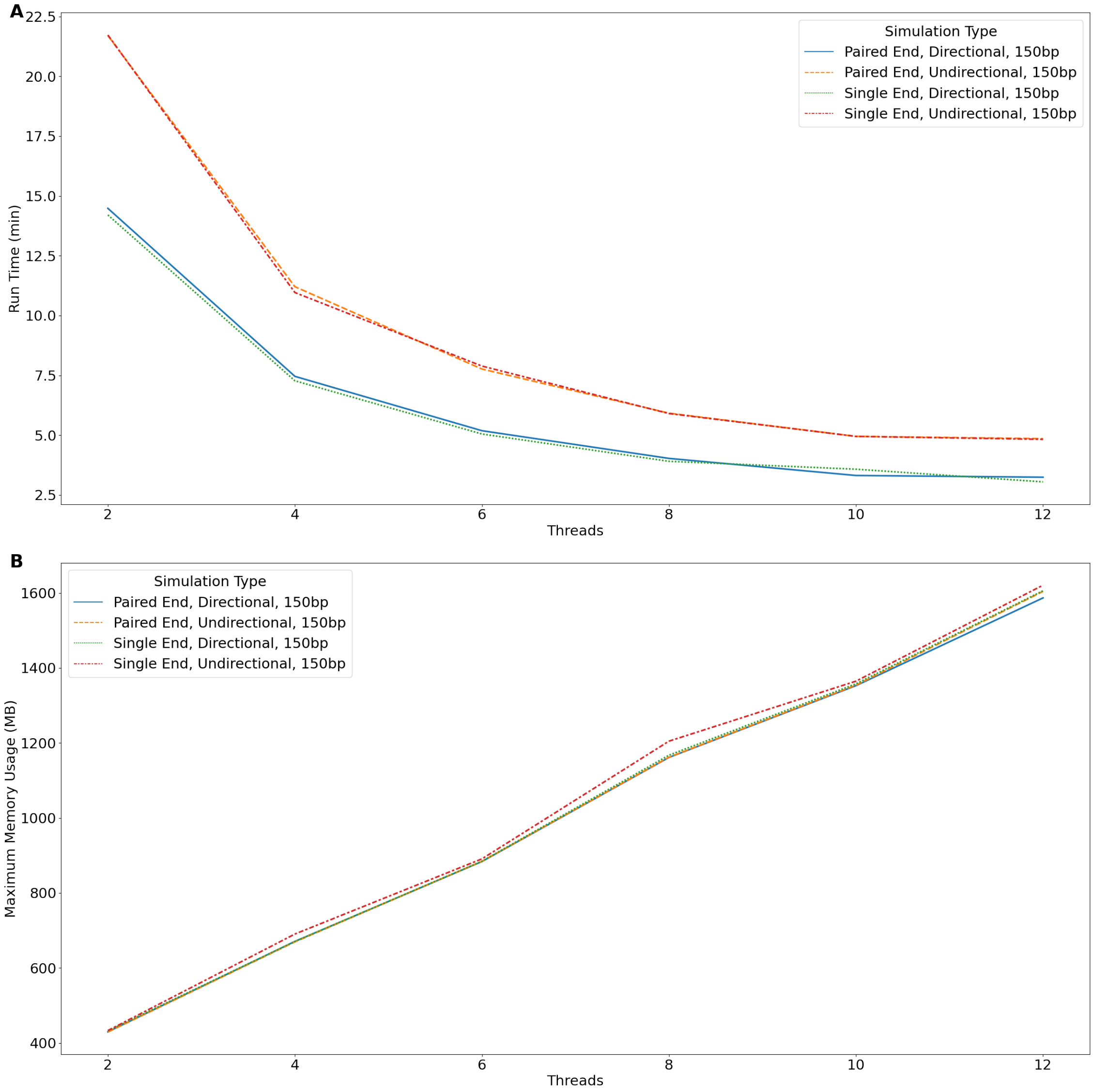
Figure 3: Targeted Bisulfite Sequencing Library Performance

(A) The number of read pairs aligned per minute for each bisulfite alignment tool. (B) Total methylation calling time (min) for each alignment file. (C) The absolute difference between array methylation values and sequencing methylation values for overlapping calls, binned by effective read depth.



*

S. Figure1: Alignment times for 150 base pair simulated libraries by the number of threads used for SAM to BAM conversion.



*

S. Figure2: (A) Total alignment time (min) and (B) Maximum memory utilization (mb) for simulated 150 bp bisulfite sequencing libraries by the number of alignment threads provided to BSBolt.

Bibliography

[1] Anaconda. *Anaconda Software Distribution.* 2020.

[2] biscuit. *biscuit.* 2021.

[3] Shawn J Cokus et al. "Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning". en. In: *Nature* 452.7184 (Mar. 2008), pp. 215–219.

[4] M Frommer et al. "A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 89.5 (Mar. 1992), pp. 1827–1831.

[5] Shicheng Guo et al. "Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA". en. In: *Nat. Genet.* 49.4 (Apr. 2017), pp. 635–642.

[6] Weilong Guo et al. "BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data". en. In: *BMC Genomics* 14 (Nov. 2013), p. 774.

[7] Steve Horvath. *DNA methylation age of human tissues and cell types.* 2013.

[8] Garrett Jenkinson et al. "Potential energy landscapes identify the information-theoretic nature of the epigenome". en. In: *Nat. Genet.* 49.5 (May 2017), pp. 719–729.

[9] Felix Krueger and Simon R Andrews. "Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications". In: *Bioinformatics* 27.11 (2011), pp. 1571–1572.

[10] Ben Langmead and Steven L Salzberg. "Fast gapped-read alignment with Bowtie 2". en. In: *Nat. Methods* 9.4 (Mar. 2012), pp. 357–359.

[11] Heng Li. "Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM". In: (Mar. 2013). arXiv: `1303.3997 [q-bio.GN]`.

[12] Heng Li. *wgsim.*

[13] Heng Li et al. “The Sequence Alignment/Map format and SAMtools”. en. In: *Bioinformatics* 25.16 (Aug. 2009), pp. 2078–2079.

[14] Wenyuan Li et al. “CancerDetector: ultrasensitive and non-invasive cancer detection at the resolution of individual reads using cell-free DNA methylation sequencing data”. en. In: *Nucleic Acids Res.* 46.15 (Sept. 2018), e89.

[15] Ryan Lister et al. “Human DNA methylomes at base resolution show widespread epigenomic differences”. en. In: *Nature* 462.7271 (Nov. 2009), pp. 315–322.

[16] Alexander Meissner et al. “Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis”. en. In: *Nucleic Acids Res.* 33.18 (Oct. 2005), pp. 5868–5877.

[17] Marco Morselli et al. “Targeted bisulfite sequencing for biomarker discovery”. en. In: *Methods* (Aug. 2020).

[18] Luz D Orozco et al. “Epigenome-wide association in adipose tissue from the METSIM cohort”. en. In: *Hum. Mol. Genet.* 27.14 (July 2018), p. 2586.

[19] Brent S Pedersen et al. “Fast and accurate alignment of long bisulfite-seq reads”. In: (Jan. 2014). arXiv: `1401.1129 [q-bio.GN]`.

[20] PyPi. *PyPI · The Python Package Index*. `https://pypi.org/`. Accessed: 2020-9-11.

[21] pysam. *pysam*. 2020.

[22] Chang Shu et al. “Comparison of methylation capture sequencing and Infinium MethylationEPIC array in peripheral blood mononuclear cells”. In: *Epigenetics & Chromatin* 13.1 (2020).

[23] Zachary D Smith and Alexander Meissner. “DNA methylation: roles in mammalian development”. In: *Nature Reviews Genetics* 14.3 (2013), pp. 204–220.

[24] A Zemach et al. “Genome-Wide Evolutionary Analysis of Eukaryotic DNA Methylation”. In: *Science* 328.5980 (2010), pp. 916–919.

[25] Michael J Ziller et al. “Charting a dynamic DNA methylation landscape of the human genome”. en. In: *Nature* 500.7463 (Aug. 2013), pp. 477–481.

CHAPTER 3

Epigenetic Suppression of Transgenic T-cell Receptor Expression via Gamma-Retroviral Vector Methylation in Adoptive Cell Transfer Therapy

Theodore S. Nowicki[1,2,3], Colin Farrell[4], Marco Morselli[5,6], Liudmilla Rubbi[6], Katie Campbell[7], Mignonette Macabali[7], Beata Berent-Maoz[7], Begoña Comin-Anduix[2,8], Matteo Pellegrini[2,5,6], and Antoni Ribas[2,3,7,8,9]

[1] Division of Pediatric Hematology-Oncology, Department of Pediatrics, University of California Los Angeles, Los Angeles, California.

[2] Jonsson Comprehensive Cancer Center, University of California Los Angeles, Los Angeles, California.

[3] Eli and Edythe Broad Center for Regenerative Medicine and Stem Cell Research, University of California Los Angeles, Los Angeles, California.

[4] Department of Human Genetics, University of California Los Angeles, Los Angeles, California.

[5] Institute for Quantitative and Computational Biosciences – The Collaboratory, University of California Los Angeles, Los Angeles, California.

[6] Department of Molecular, Cell, and Developmental Biology, University of California Los Angeles, Los Angeles, California.

[7] Division of Hematology-Oncology, Department of Medicine, University of California Los Angeles, Los Angeles, California.

[8] Division of Surgical Oncology, Department of Surgery, University of California Los Angeles, Los Angeles, California.

[9] Department of Molecular and Medical Pharmacology, University of California Los Angeles, Los Angeles, California.

Transgenic T-cell receptor (TCR) adoptive cell therapies recognizing tumor antigens are associated with robust initial response rates, but frequent disease relapse. This usually occurs in the setting of poor long-term persistence of cells expressing the transgenic TCR, generated using murine stem cell virus (MSCV) γ-retroviral vectors. Analysis of clinical transgenic adoptive cell therapy products in vivo revealed that despite strong persistence of the transgenic TCR DNA sequence over time, its expression was profoundly decreased over time at the RNA and protein levels. Patients with the greatest degrees of expression suppression displayed significant increases in DNA methylation over time within the MSCV promoter region, as well as progressive increases in DNA methylation within the entire MSCV vector over time. These increases in vector methylation occurred independently of its integration site within the host genomes. These results have significant implications for the design of future viral-vector gene engineered adoptive cell transfer therapies.

3.1 Introduction

Genetically engineered adoptive cell therapy (ACT) is revolutionizing cancer treatment, with sustained clinical responses seen in a variety of malignancies. Current approaches utilize retroviral or lentiviral vectors for ex vivo transduction of a patient's T-cells to express either a cancer antigen-specific T-cell receptor (TCR) or a chimeric antigen receptor (CAR). These reinfused cells then create a focused anti-tumor response in a variety of cancer subtypes [25, 12]. However, while these treatments lead to durable clinical responses in many patients, a significant number of patients remain who do not respond, or who eventually relapse.

Previous ACT clinical trials conducted by our group and others against the tumor antigens MART-1 (in melanoma) and NY-ESO-1 (in sarcoma and melanoma) have demonstrated that detectable surface expression of the transgenic TCR is rapidly lost in circulating T-cells following infusion [17, 3, 18, 21]. The transduction of these cells relies on retroviral vectors, most commonly the murine stem cell virus (MSCV), a γ-retrovirus which has been optimized for highly efficient transgene expression, and has been used for a variety of such applications in vivo [10]. However, it has subsequently been shown to be vulnerable to epigenetic silencing via DNA methylation of CpG loci which are clustered within its 5' long tandem repeat (LTR) promoter region [23, 27].

Given our observation of this phenomenon of rapid loss of surface expression of the transgenic TCR in circulating T-cells in vivo, along with the vulnerability of the MSCV vector to epigenetic silencing via DNA methylation, we hypothesized that acquisition of DNA methylation within the retroviral 5'LTR promoter was associated with loss of expression of the transgenic TCRs in these clinical samples. Herein we describe the analysis of clinical transgenic ACT samples for persistence of the transgenic TCR DNA sequence and accompanying expression of the TCR itself, as well as characterizing the DNA methylation status of the MSCV vector over time, and the relationship between vector methylation and suppression of transgenic TCR expression.

3.2 Results

Trial conduct, patient characteristics, and outcomes

16 patients from our previous transgenic TCR ACT trials directed against MART-1 [3] and NY-ESO-1 [18] were selected for analysis. Patient demographics, clinical characteristics, and outcomes are summarized in Table 1. Following conditioning chemotherapy, patients were all treated with up to $1x10^9$ autologous transgenic TCR T-cells, which were generated via ex vivo transduction using the MSCV γ-retrovirus encoding for the F5-MART-1 TCR or

the NY-ESO-1 TCR (Figure S1). Of the 16 patients selected, seven out of eight patients treated with F-5 MART-1 TCR transgenic T-cells, and six out of eight patients treated with NY-ESO-1 TCR transgenic T-cells, demonstrated a transient objective response to therapy.

Transgenic TCR-engineered T-cells display strong persistence of the transgene DNA sequence, but with greatly reduced expression of the RNA and surface protein over time

The above 16 patients had peripheral mononuclear blood cell (PBMC) samples from both their infusion (day 0) and 70 days after treatment (in peripheral circulation) analyzed for persistence of the transgenic TCR and RNA and surface TCR protein expression. All infusion products demonstrated robust presence of the transgenic TCR gene and its RNA and surface protein expression. However, we observed that despite largely decreased expression of the RNA transcript and the surface protein at day +70, persistence of the transgenic TCR still accounted for the vast majority of circulating TCR DNA clonotypes as measured by TCR sequencing (Figure 1A, Figure S1, S2, S3). Six of these patients with surface TCR protein expression <0.5% of circulating CD3+ cells, the established threshold of a highly expanded clone (HEC) [13], were designated as an "expression-low" cohort for further analyses, while the remaining ten patients were designated as "expression-high." While the degree of RNA and surface protein expression of the transgenic TCRs were significantly lower at day +70 in the expression-low group compared to the expression-high group, there were no statistically significant differences between the two cohorts' proportion of transgenic TCR DNA in circulating PBMCs at day +70, which still accounted for the majority of circulating TCR clonotypes (Figure 1B-D, Figure S4, S5). There were also no statistically significant differences between the two groups' transgenic TCR proportions and surface protein expression at day 0 (Figure S6).

Increased MSCV 5'LTR methylation is associated with decreased expression of the transgenic RNA and protein, despite persistence of the transgenic DNA

Given the overall strong predominance in transgenic TCR's representation within the TCR repertoire of circulating T-cells despite profound loss of its expression, we explored the degree of CpG methylation within the MSCV 5' LTR promoter region, which contains a CpG island characterized by a high concentration of clustered CpG loci over a relatively small region of DNA (Figure 2A). Genomic DNA was isolated from PBMCs of each patient sample at baseline (day 0) and 70 days post-infusion, and bisulfite converted. The area of DNA within the MSCV 5'LTR containing the CpG island was amplified by PCR, purified, and sequenced. We found that while all day 0 infusion products contained relatively little CpG methylation within the 5'LTR promoter region of the MSCV vector, the six patients in the expression-low cohort individually demonstrated significantly increased levels of CpG methylation at day +70 (Figure 2B-C, Figures S7, S8). The average proportion of promoter methylation was anticorrelated with transgenic TCR surface protein expression among all patients at day +70 (Figure S9). Furthermore, when the two cohorts were compared with one another, the expression-low cohort displayed significantly greater CpG methylation within the 5'LTR promoter region when compared to the expression-high cohort at day +70 (Figure 2D). While there were no differences in progression-free survival or overall survival between the two cohorts, the expression-low cohort displayed inferior decrease in tumor burden compared to the expression-high cohort, which nearly achieved statistical significance ($p = 0.07$, Figure S10).

CpG methylation is increased across MSCV vector over time in all patients, and is significantly greater in those with decreased transgenic TCR expression over time

In order to expand our ability to characterize CpG methylation across the entire MSCV vector over time, we performed bisulfite conversion on genomic DNA library preparations isolated from all patients' PBMC samples at day 0 (infusion), day +30, and day +70. We then carried out target enrichment using RNA probes to capture the MSCV vector. Bisulfite-converted libraries were then aligned against the human genome version 38 (hg38) with MSCV transgenic TCR vector reference sequences utilizing BSBolt, an integrated alignment and analysis platform for bisulfite-converted DNA. Each patient sample demonstrated overall progressive increases in CpG methylation across the transgenic TCR MSCV vector sequence over time (Figure 3A-B). When all patient sample data were aggregated, the increases in MCV vector CpG methylation were statistically significant at day +30 compared to day 0, and day +70 compared to day +30 (Figure 3C). When the data were further stratified to compare the expression-high and expression-low patient cohorts, we observed that while there were no significant differences between baseline levels of CpG methylation at day 0, the day +30 and day +70 methylation ratios, which progressively increased over time in both cohorts, were significantly higher in the expression-low cohort when compared to the expression-high cohort, consistent with what was observed in the targeted analysis of the 5'LTR promoter region (Figure 3D).

MSCV-TCR vector integration occurs sporadically throughout the host genome relative to transcription start sites

While capturing the MSV vector fragments, we also obtained fragments that spanned the junction of the insertion site between the vector and the genome. These reads allowed us to explore the integration patterns of the MSCV vector within the patients' genomes.

Specifically, we utilized BSBolt to map discordant paired reads, where individual reads from the pair map to both the human genome and the MSCV transgenic TCR vector. Integration sites were characterized by their relative and absolute distance to transcription start sites (TSS). We observed that the MSCV vector integration sites occurred at sites generally distal to gene TSS (Figure S11). Furthermore, when we compared the proportions of vector integration sites by their relative distance to TSS, we observed no significant differences between the expression-low cohort (i.e. those with high CpG methylation levels) and the expression-high cohort (i.e. those with low CpG methylation levels), suggesting that the integration site relative to TSS did not impact the degree of CpG methylation observed within a given vector read.

3.3 Discussion

TCR transgenic ACT has established itself as a potent form of cancer immunotherapy for a wide variety of tumor subtypes. However, despite frequent early responses and reduction in tumor burden, the durability of these responses is often poor, and tumors often progress within several months. Our clinical experiences with transgenic TCR ACT generated with γ-retroviral vectors, as well as those of other groups, have consistently demonstrated that the presence of detectable circulating transgenic TCR surface expression rapidly diminishes within 1-2 months following cell transfer, in keeping with the timeline of disease progression after the initial transient response to therapy [17, 3, 18, 21]. This is in stark contrast to ACT using autologous cancer-antigen-specific TCR clones, which are isolated, expanded ex vivo, and reinfused to the patient without the aid of any viral vectors to transduce and generate these cells in large numbers. Previous ACT studies utilizing such endogenous TCR clones have shown remarkably strong persistence of cancer-antigen-specific TCR clones in circulation following cell transfer [8, 11]. This discrepancy implies that genetically engineered ACT products have a fundamental vulnerability in the suppression of their transgenic TCR.

Given the known vulnerability of the MSCV vector to epigenetic silencing via CpG methylation, as well as the observed discordance between the strong persistence of the TCR transgenes and their poor expression in circulation, we hypothesized that increases in DNA methylation within the vector were associated with this phenomenon. While one small series of patients previously found no significant retroviral promoter methylation to cause suppression of transgenic TCR expression [2], that study only examined the methylation status of the MSCV vector promoter in a total of four patients. Our examination of 16 samples from patients receiving transgenic ACT demonstrated that samples from only six of these patients displayed significantly discordant, profound decreases in expression of the transgenic TCR which was associated with significant increases in MSCV vector methylation. This suggests that the phenomenon occurs in a minority of patients, and could be missed by sampling too small a cohort. Indeed, there may be cell phenotype-specific or patient-specific predispositions to rapid acquisition of CpG methylation of γ-retroviruses which would not be present in every subject studied. While further detailed studies of factors such as patient-specific polymorphisms in DNA methyltransferase enzymes would be needed to derive even speculative inferences into such predispositions, our studies did not demonstrate any significant association with the MSCV integration site's distance to a given TSS at infusion and its propensity to acquire CpG methylation over time. Furthermore, our characterization of the MSCV vector integration sites was consistent with previously published studies dealing with the γ-retroviral vector murine leukemia virus (MLV), which showed that only 25% of γ-retrovirus integration sites are within ±2.5-kb around the TSS [7] in human cells, consistent with our results. While MSCV has previously shown increased integration near TSS in murine bone marrow cells [1], it may be that there are species-specific factors which influence integration site, as our data are consistent with previously published data in humans.

The vulnerability of retroviral vectors to epigenetic suppression in transgenic TCR ACT products seen here raises an important question about how to overcome this potential weakness in clinical practice. One possibility would be to utilize dual therapy with systemic hypomethylating agents such as decitabine to prevent the acquisition of CpG methylation within the retroviral vector encoding the transgenic TCR. Such agents have also been shown in murine models to increase the expression of tumor antigens commonly targeted by transgenic TCR ACT, such as NY-ESO-1 [4, 20], theoretically further enhancing the immunotherapeutic effect of the transgenic T-cells. However, DNA methylation is often followed by histone recruitment and modification (acetylation and/or methylation), which further contribute to epigenetic suppression. Therefore, such pharmacologic interventions would potentially be insufficient to fully reverse epigenetic suppression, as it would not have an effect on the histone modifications and recruitment. Stimulation of T-cells with IL-2 and anti-CD3/CD28 beads has previously been shown to partially restore transgenic TCR expression in vitro in some patients, likely due to nonspecific modulation of these epigenetic factors [2]. New non-viral approaches to generating transgenic TCR T-cells using CRISPR-Cas9 to deliver constructs under control of the native TCR promoter (rather than a viral promoter) would also potentially avoid this risk of epigenetic suppression of viral vector-encoded products [22]. Furthermore, there may be utility in new modalities that provide a continuous supply of transgenic T-cells to the patient. Preclinical models have demonstrated that CD34+ hematopoietic stem cells encoding a transgenic TCR can endogenously differentiate into fully functional T-cells expressing the TCR [26, 24]. We currently have recently an open phase I clinical trial which utilize this approach against NY-ESO-1 in solid tumors (NCT03240861), utilizing a lentiviral vector for stem cell transduction for long-term expression.

One major limitation of our study is that we only examined transgenic TCR ACT products generated using the MSCV γ-retrovirus, which is known to be potentially vulnerable to epigenetic silencing via CpG methylation. Other types of cell therapy products, including the CD19 CAR-T product Kymriah^{TM} (tisagenlecleucel), utilize lentiviral vectors, which have previously demonstrated remarkable persistence of transgene expression in vivo and do not appear vulnerable to CpG methylation [19], which may limit the broader applicability of our findings. Indeed, transgenic TCR ACT products manufactured using lentiviral vectors have been shown to have far superior persistence of detectable surface expression of the TCR [6]. While the other commercially available CD19 CAR-T product Yescarta™ (axicabtagene ciloleucel) does utilize a γ-retroviral vector for its manufacture, its rates of durable complete and partial remission are far superior to any published transgenic TCR product [12]. This is likely due to the rapid systemic clearance of lymphoma cells seen when using these ACT products, implying that long-term persistence of the transgenic T-cells in this setting is potentially less important than in treating refractory solid tumors with such therapeutics. Indeed, our examination of patient samples at day +70 was chosen due to this being the average point of circulating transgenic T-cell nadir in our previously published trials [3, 18]. We were unable to determine any association of γ-retroviral vector methylation with patient survival, likely owing to the overall small number of long-term responders inherent to this therapy in solid tumors.

In summary, we have shown that progressive increases in CpG methylation within the MSCV γ-retroviral vector are associated with rapid suppression of transgenic TCR expression over time in clinical transgenic ACT, despite strong persistence of the transgene itself. This phenomenon did not appear to have any correlation with the vector integration site within the host genome. These findings have significant implications in how the cellular therapeutics community should approach the design of future generations of these products.

3.4 Materials and Methods

Clinical trial, patients, and manufacturing of MART-1 and NY-ESO-1 TCR engineered T-cells

For the F5-MART-1 transgenic TCR adoptive cell therapy clinical trial, patients positive for HLA-A*0201 with a MART-1-positive metastatic melanoma were enrolled under NCT00910650 (UCLA IRB #08-02020 and #10-001212) from April 2009 to September 2011, under investigational new drug (IND) #13859 [3]. For the NYESO-1 transgenic TCR adoptive cell therapy clinical trials, patients positive for HLA-A*0201 with an NYESO-1-positive sarcoma or melanoma were enrolled under NCT02070406 or NCT01697527 (UCLA IRB #12-000153 and #13-001624, respectively) under IND#15167 [18]. Clinical trial design and manufacturing of the MART-1 and NYESO-1 TCR transgenic T-cells are previously described [3, 18]. Briefly, non-mobilized autologous PBMCs were stimulated in culture with IL-2/OKT3 and transduced with clinical grade MSCV retrovirus vector expressing the MART-1 F5 TCR or the NYESO-1 TCR on two consecutive days, then continually expanded ex vivo for 6-7 days. Up to $1x10^9$ transgenic TCR transgenic lymphocytes were administered to each patient following conditioning chemotherapy with cyclophosphamide and fludarabine, along with post-infusion systemic IL-2 for 7-14 days, and d endritic cell vaccine boosts, as previously described [3, 18].

Quantification of transgenic TCRβ genomic DNA persistence

Genomic DNA and RNA was isolated from patient-matched infusion products and post-infusion PBMCs recovered at day +70 (+/- 10 days), with an AllPrep DNA/RNA isolation kit according to the manufacturer's instructions (Qiagen). TCRβ alleles were sequenced at 100,000 reads by Adaptive Biotechnologies (Seattle, WA). Briefly, this process utilizes a synthetic immune repertoire, corresponding to every possible biological combination of Variable (V) and Joining (J) gene segments for each T-cell receptor locus, spiked into every sample at a known concentration. These inline controls enable rigorous quality assurance for every sample assayed and allow for correction of multiplex PCR amplification bias, providing an absolute quantitative measure of T-cells containing the transgenic TCR relative to the other endogenous TCR clonotypes, with no difference in amplification efficiency [6]. Productive TCRβ sequences, i.e. those that could be translated into open reading frames and did not contain a stop codon, were reported. The transgenic F5-MART-1 and NY-ESO-1 TCR sequences' persistence were identified based on comparison of reads with the known TCRβ

sequence for the transgenic product, and expressed as a percentage of total productive TCRβ sequences present within a given sample/timepoint.

qRT-PCR

Total RNA isolated from patient samples (as described above) was used for analysis of relative abundance of the transgenic F5 MART-1 TCR or the NYESO-1 TCR. Samples were converted to cDNA using iScript TM Reverse Transcription Supermix for RT-PCR (Bio-Rad), then cDNA was amplified and quantified using iTaq TM Universal SYBR Green Supermix (Bio-Rad) on an Applied Biosystems 7500 Fast Real-Time PCR System (Applied Biosystems). PCR conditions were 1 cycle of 1 min at 95°C, 35 cycles of 15 sec at 95°C and 60 sec at 60°C, and 5 min incubation at 72°C. Replicate samples were run with test primer sets for the F5-MART-1 TCR, the NYESO-1 TCR, or the endogenous control glyceraldehyde-3-phosphate dehydrogenase (GAPDH); primer sequences are available upon request. Data were analyzed according to the comparative Ct method.

MHC dextramer immunologic monitoring for surface expression of transgenic TCRs

Detection and quantification of F5 MART-1 TCR or NY-ESO-1 TCR expression using fluorescent MHC dextramer analysis for MART-1 or NY-ESO-1 (Immudex) was performed on patient-matched infusion products and post-infusion PBMCs recovered at day +70, as previously described [3, 18, 5]. Our definitions for a positive or negative immunologic response using standardized MHC multimer assays were used, which are based on assay performance specifications by defining changes beyond the assay variability with a 95% confidence level [5].

Bisulfite sequencing of MSCV retroviral promoter in patient samples

Genomic DNA isolated from patient PBMC samples was bisulfite converted using the EZ DNA Methylation-Gold Kit (Zymo Research) according to the manufacturer's instructions. A CpG island within the MSCV 5'LTR promoter U3/R/U5 region, defined as an area >100bp, with a GC content of >50%, and possessing an observed:expected CpG ratio of >0.6, was determined using MethPrimer software [16], which also designed PCR primers capable of amplifying the methylated and non-methylated bisulfite-converted DNA sequence of interest. CpG islands were PCR amplified, purified, and subjected to DNA Sanger sequencing (Laragen). The methylation status of each CpG locus within the individual amplicons was determined using QUMA software [14].

Targeted bisuflite sequencing library preparation and sequencing

Purified genomic DNA from patient samples (isolated as described above) was quantified using the Qubit dsDNA BR Assay (Thermo Fisher Scientific). For each sample, 250ng of DNA was sonicated using a Bioruptor Pico (Diagenode) for 15 cycles (30 sec ON; 60 sec OFF). Libraries were prepared using the NEBNext Ultra II DNA kit (NEB) according to manufacturer instructions with few modifications. Briefly, sonicated DNA was subjected to EndPrep (End Repair and A-tailing), followed by Adapter Ligation using 2.5µL of Illumina TruSeq pre-methylated Adapters (Illumina). Samples were purified using 0.85x NEB Purification Beads and eluted in 15µL of 10mM Tris-HCl, pH 8. Samples were mixed in 16-sample pools and column purified using a DCC-5 (Zymo Research). Elution was performed with 10µL of 60℃ 10mM Tris-HCl, pH 8. Each sample pool was subject to hybrid capture with custom biotinylated RNA probes designed to tile the MSCV vector (MyBaits Arbor Bioscience - Human_6K and Human patch2) according to the manufacturer's protocol. The hybridization was carried out for 20 hours overnight at 65℃. Captured DNA was eluted by heating in 20µL of 10mM Tris-HCl pH 8+0.05% Tween-20. The eluted DNA was then subject to bisulfite conversion using the DNA Methylation Lightning kit (Zymo Research). Converted DNA was then amplified using xGen Library Amplification Primer Mix (IDT) and Kapa Uracil+ Ready Mix using the following conditions: 98°C for 2 min; 20 cycles of 98°C-20 sec, 60°C-30 sec, 72°C-30 sec; Final Extension 72°C-5 min; hold 4°C. PCR products were purified using 0.9 volumes of NEBNext Purification Beads and eluted in 15µL of 10mM Tris-HCl, pH 8. Final libraries were then quantified using the Qubit dsDNA BR Assay (Thermo Fisher Scientific) and visualized using a D1000 ScreenTape (TapeStation 2200 system - Agilent Technologies). Each pool was then sequenced at 150bp PE on a HiSeq3000 instrument (Illumina).

Targeted Bisulfite Sequencing Alignment and Methylation Calling

Paired end, 150bp targeted bisulfite sequencing reads were aligned to the combined hg38 and MSCV transgenic TCR vector sequence bisulfite converted references using BSBolt v0.1.2 [9] local alignment. Alignments with ¡5 mismatches and an alignment score ¿160 were considered valid, up to 10 alignments per read pair were considered. Alignments where read pairs did not meet expected paired end constraints, an insert size ¿500bp or alignments on separate chromosomes, were reported as discordant. Reads pairs with only one valid alignment were reported as mixed. Duplicate reads were removed using SAMtools v1.9 [15]. Following duplicate removal methylation values were called for all observed cytosines with $\geq$5 reads with a base call quality above 25 using BSBolt v0.1.2 [9].

Vector Integration Site Detection

Discordant reads pairs with a vector alignment and a genome alignment were evaluated as potential integration sites. Discordant read pairs were further filtered by removing reads that aligned to genomic regions homologous with the vector sequence or aligned outside the expected integration region within the vector sequence. Alignments with an alignment score greater than 160 and with secondary alignments that repeated no more than 10% of the primary alignment sequence were considered integration site supporting alignments. Integration sites were reported as the closest genomic base to vector alignment or the average integration site position for sites with multiple supporting reads. The integration site selection pipeline was implemented using custom python code (https://github.com/NuttyLogic/Epigenetic-suppression-of-transgenic-TCR-expression-in-ACT). The vector integration detection pipeline was validated against simulated 150bp, paired end bisulfite converted vector integration libraries; see Supplemental Methods for further details.

Statistical analysis

Graphing and descriptive statistical analyses were carried out with GraphPad Prism version 7.0 (GraphPad). Where indicated, Mann-Whitney U test or Wilcoxon matched-pairs signed rank test were used for comparison of two groups, and correlations between CpG promoter methylation and transgenic TCR expression were compared using Spearman rank correlation. P values of ¡0.05 were considered statistically significant.

Corresponding author: Theodore S. Nowicki, M.D., Ph.D.; Jonsson Comprehensive Cancer Center (JCCC) at the University of California Los Angeles (UCLA), 12-159 Factor Building, 10833 Le Conte Avenue, Los Angeles, CA, 90095. Phone: 310-267-5145; Fax: 310-825-2493; Email: tnowicki@mednet.ucla.edu.

GRANT SUPPORT: This study was funded in part by NIH grants R35 CA197633 and P01 CA168585, the Parker Institute for Cancer Immunotherapy, and the Ressler Family Fund (to A.R.). T.S.N. is supported by the NIH/NICHD grant K12-HD000850 (Pediatric Scientist Development Program), the Tower Cancer Research Foundation Career Development Award, and the Hyundai Hope on Wheels Young Investigator Award.

DISCLOSURE OF POTENTIAL CONFLICTS OF INTEREST: T.S.N. has received honoraria from consulting with Allogene Therapeutics. A.R. has received honoraria from consulting with Amgen, Bristol-Myers Squibb, Chugai, Genentech, Merck, Novartis, Roche and Sanofi, is or has been a member of the scientific advisory board and holds stock in Advaxis, Arcus Biosciences, Bioncotech Therapeutics, Compugen, CytomX, Five Prime, FLX-Bio, ImaginAb, Isoplexis, Kite-Gilead, Lutris Pharma, Merus, PACT Pharma, Rgenix and Tango Therapeutics, and has received research funding from Agilent and Bristol-Myers Squibb through Stand Up to Cancer (SU2C). B.C-A. has received honoraria from consulting with Advarra IBC. The rest of the authors declare no potential conflicts of interest.

DISCLAIMER: The contents of this article are solely the responsibility of the authors and do not necessarily represent the official view of the NIH, the NCI, the NICHD, the Parker Institute for Cancer Immunotherapy, or the Ressler Family Foundation.

Supplemental Methods

Insertion Site Search Validation

Vector insertion events were simulated by randomly sampling positions from the human genome (hg38), excluding alt contigs and sex chromosomes, between 20 - 30 times per simulation library. Bisulfite sequencing reads (125bp, paired-end) were then simulated from the vector insertion contigs and mapped to a combined hg38 and MSCV transgenic TCR reference (BSBolt v0.1.2). Six simulation libraries, three MART-1 TCR and three NY-ESO-1 TCR, were created. The vector insertion detection pipeline was then validated against the known integration positions from the simulated reads. Correct identification of alignments that spanned the genome or vector sequence as a split (1 or more spanning bases) or discordant (complete read end mapped) was assessed for each library for various minimum

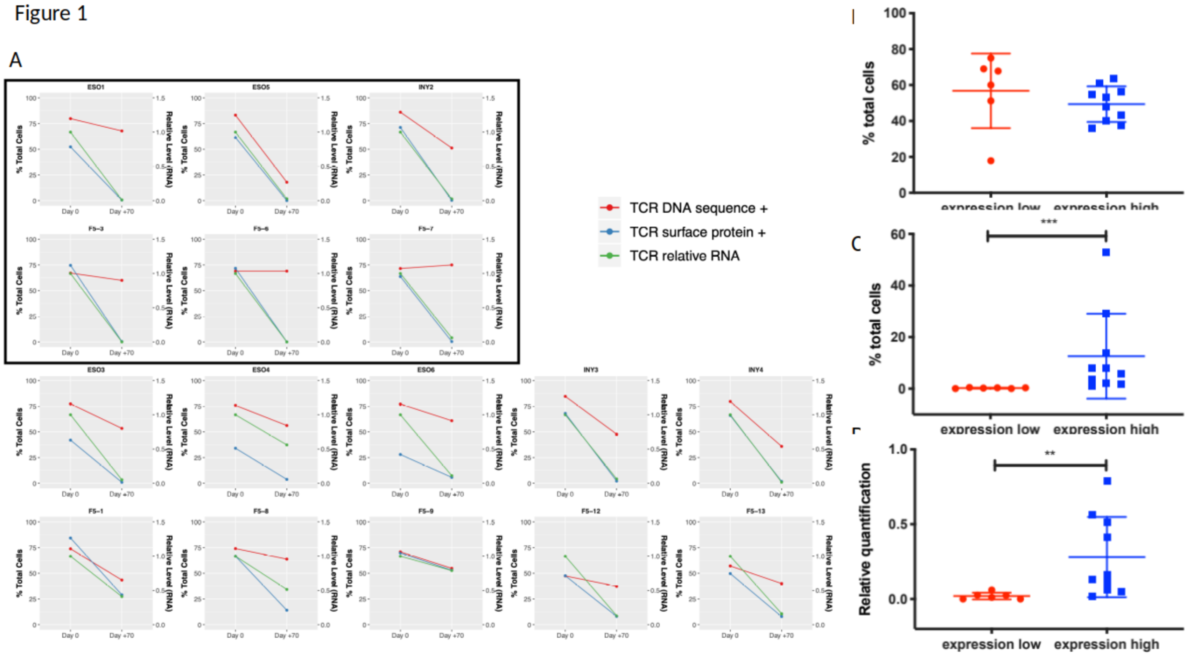
*

Table 1: Patient Demographics and Outcomes

Abbreviations: F: female; LN: lymph nodes; M: male; PFS: progression-free survival; OS: overall survival; EOS: end of study; DC: dendritic cells; TCR: T cell receptor; IL-2: interleukin-2; mo: months

Patient study number	**Sex**	**Ethnicity**	**Age**	**Type of Cancer**	**Active Disease Sites**	**Stage**	**Number of TCR transgenic cells**	**IL-2 doses**	**DC doses**	**Evidence of transient tumor response**	**Response at EOS (day 90)**	**PFS (mo)**	**OS (mo)**	**Current Status**
F5-1	M	Caucasian	60	Melanoma	Lung, Stomach, Liver, Pancreas, Peritoneum, Soft tissues	M1c	$1x10^9$	12/14/21	03/03/21	Yes by PET/CT	PD	3.0	5.0	Died of disease
F5-3	M	Caucasian	61	Melanoma	Lung, Liver	M1c	$1x10^9$	13/14	03/03/21	Yes by PET/CT	SD	7.0	86.0	Died of disease
F5-6	M	Caucasian	59	Melanoma	Lung, LN	M1b	$1x10^9$	13/14	03/03/21	Yes by PE	SD	3.0	4.0	Died of disease
F5-7	M	Caucasian	48	Melanoma	SC, Bone	M1c	$1x10^9$	09/14/21	03/03/21	Yes by CT	SD	4.0	11.0	Died of disease
F5-8	M	Caucasian	44	Melanoma	LN, Liver	M1c	$1x10^9$	11/14/21	03/03/21	Yes by PET/CT	SD	4.0	11.0	Died of disease
F5-9	F	Caucasian	46	Melanoma	Skin, LN	M1a	$1x10^9$	11/14/21	03/03/21	No	PD	3.0	20.0	Died of disease
F5-12	M	Caucasian	40	Melanoma	Lung, LN	MIVb	$3.9x10^9$	06/09/21	03/03/21	Yes by PET/CT	SD	5.0	8.0	Died of disease
F5-13	M	Caucasian	60	Melanoma	Lung, Abdomen, SC	MIIIb	$4.41x10^9$	04/09/21	03/03/21	Yes by PET/CT	SD	3.0	8.0	Died of disease
ESO-1	M	Hispanic	47	Liposarcoma	Right Renal Fossa; Liver left Lobe; Hepatic Segment; Peritoneal, Perihepatic Nodule	IV	$7.7x10^8$	28/28	03/03/21	No	PD	2.6	16.0	Died of disease
ESO-3	F	Caucasian	24	Synovial Sarcoma	Right Lung; Multiple pulmonary Nodules	IV	$1x10^9$	19/28	01/03/21	Yes by PET/CT	PR	67.0	67.0	Alive with CR
ESO-4	M	Caucasian	41	Synovial Sarcoma	Left Infraclavicular Mass; Left Pectoralis Mass	III	$1x10^9$	18/28	03/03/21	Yes by PET/CT	PD	3.0	25.0	Died of disease
ESO-5	F	Caucasian	43	Synovial Sarcoma	Right popliteal fossa; Lung	IV	$1x10^9$	14/14	03/03/21	Yes by PET/CT	PR	9.0	41.5	Died of disease
ESO-6	M	Caucasian	26	Osteosarcoma	Lung	IV	$1x10^9$	14/14	03/03/21	Yes by PET/CT	PD	2.5	19.0	Died of disease
INY-2	M	Caucasian	66	Melanoma	LN, Liver	IV	$1x10^9$	21/28	02/03/21	Yes by PET/CT	PD	3.0	6.0	Died of disease
INY-3	F	Caucasian	44	Synovial Sarcoma	Right popliteal fossa; Lung	IV	$1x10^9$	14/14	03/03/21	Yes by PET/CT	PD	3.0	31.0	Died of disease
INY-4	M	Hispanic	24	Melanoma	Lung, LN, adrenal gland, liver, trachea, brain	IV	$1x10^9$	10/14/21	03/03/21	No	PD	NaN	3.0	Died of disease

Figure 1

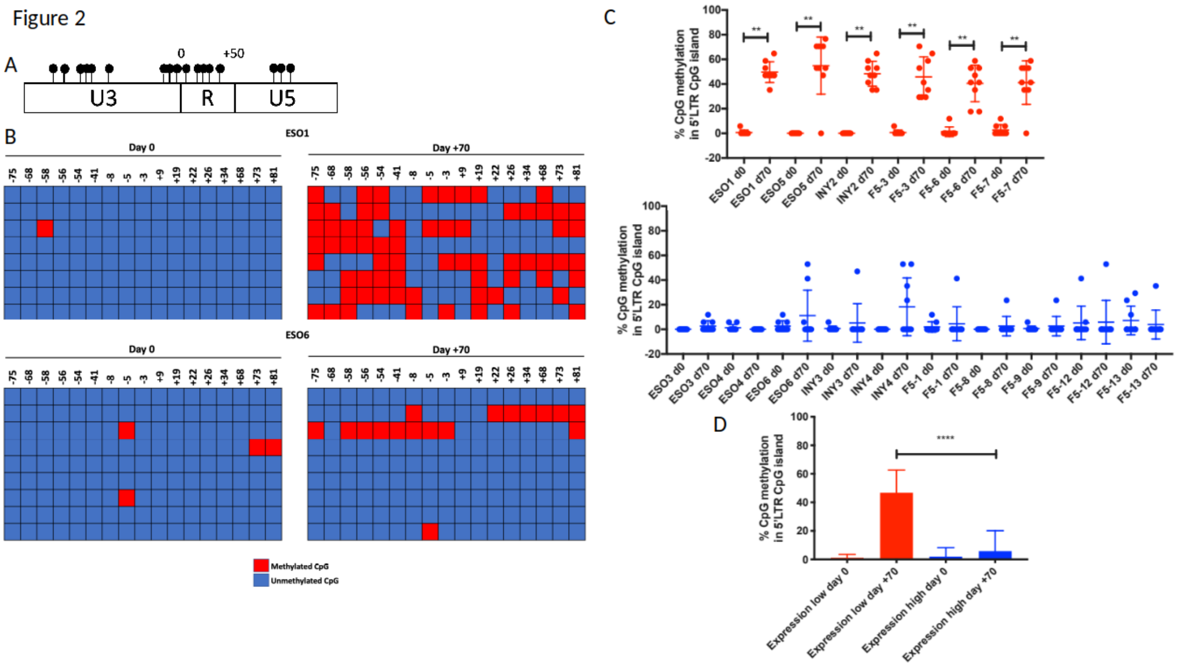


*

Figure1: Persistence of transgenic TCR DNA and RNA/protein expression of TCR-engineered T-cells over time

A) Comparison between infusion products (day 0) and post-infusion recovery products 70 days later from patients on NYESO-1 TCR-engineered cell therapy trials (ESO and INY) or F5-MART-1 TCR-engineered cell therapy trial (F5). Data displayed are percentage of cells containing the transgenic TCR DNA sequence (red, left axis), the percentage of cells expressing the TCR protein (green, left axis), and the relative level of RNA transcript (blue, right axis). Inset boxed patients represent those with surface TCR expression ¡0.5% at day +70. B, C, D) Statistical comparisons between expression-high and expression-low patient cohorts demonstrating no significant differences between day +70 transgenic TCR DNA (B), while surface protein expression (C) and relative RNA level (D) at day +70 are significantly lower in the expression-low cohort (** $p < 0.01$, ***$p < 0.001$, Mann-Whitney U test).

Figure 2

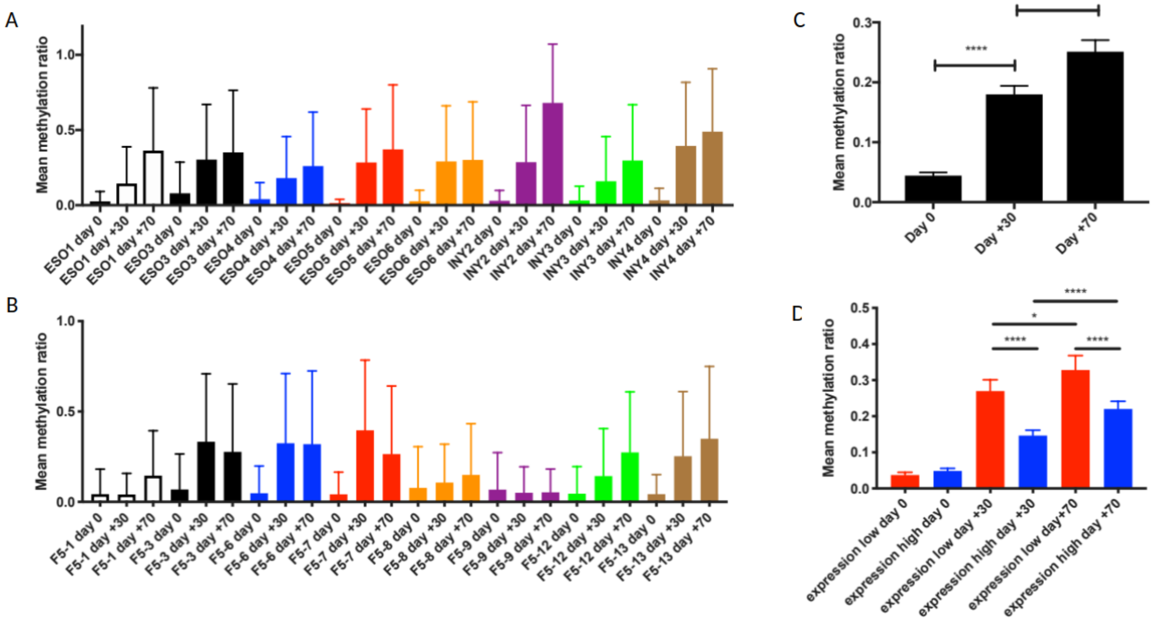


*

Figure2: Increased MSCV 5'LTR methylation is associated with decreased expression of the transgenic RNA and protein, despite persistence of the transgenic DNA.

A) CpG island within the MSCV 5' LTR, where each CpG loci is represented by a filled circle. Numbering is relative to the transcription start site. B) Bisulfite conversion was performed on genomic DNA from patient PBMCs at day 0 (infusion) and day +70, and the CpG island within the MSCV 5'LTR promoter region was PCR amplified and purified; two representative patients are shown. Each row represents sequencing of an individual experiment, with methylated cytosine loci indicated by red boxes and unmethylated loci by blue boxes. CpG loci positions are listed at the top of the graph relative to the TSS for the transgenic TCR. C) Percentage of CpG methylation within 5'LTR at day 0 and day +70 in individual expression-low (red) and expression-high (blue) patients (** $p < 0.01$, Wilcoxon matched-pairs signed rank test); comparison between aggregate percent CpG methylation within the 5'LTR between all expression-high and expression-low patients at day 0 and day +70 is shown in D (**** $p < 0.0001$, Mann-Whitney U test).

Figure 3

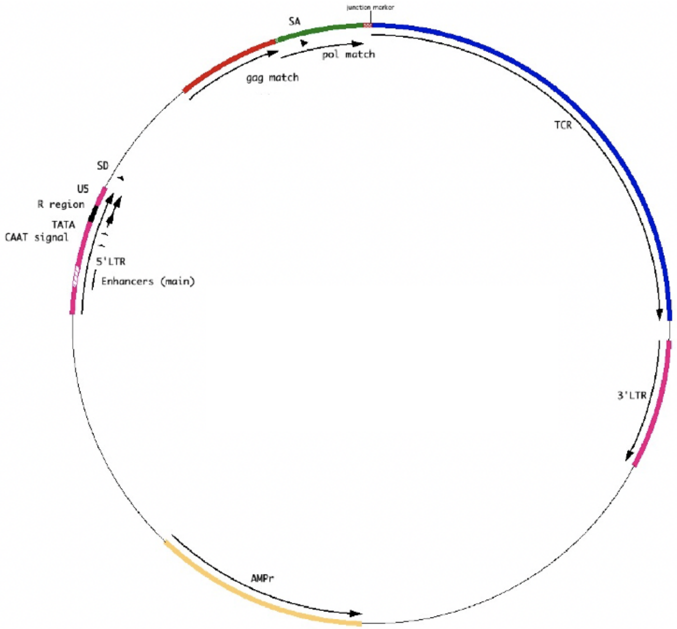


*

Figure3: CpG methylation is increased across MSCV vector over time in all patients, and is significantly greater in those with decreased transgenic TCR expression over time.

Mean methylation ratio across MSCV vector at day 0, day +30, and day +70 for each patient treated with NY-ESO-1 TCR (A) and F5-MART-1 TCR (B) transgenic T-cells. C) Statistical comparisons between methylation values in all patients at day 0, day +30, and day +70; data are stratified to compare the increases in methylation values over time between the expression-high and expression-low patients in D (*, p¡0.05, **** p¡0.0001, Mann-Whitney U test).

Figure S1

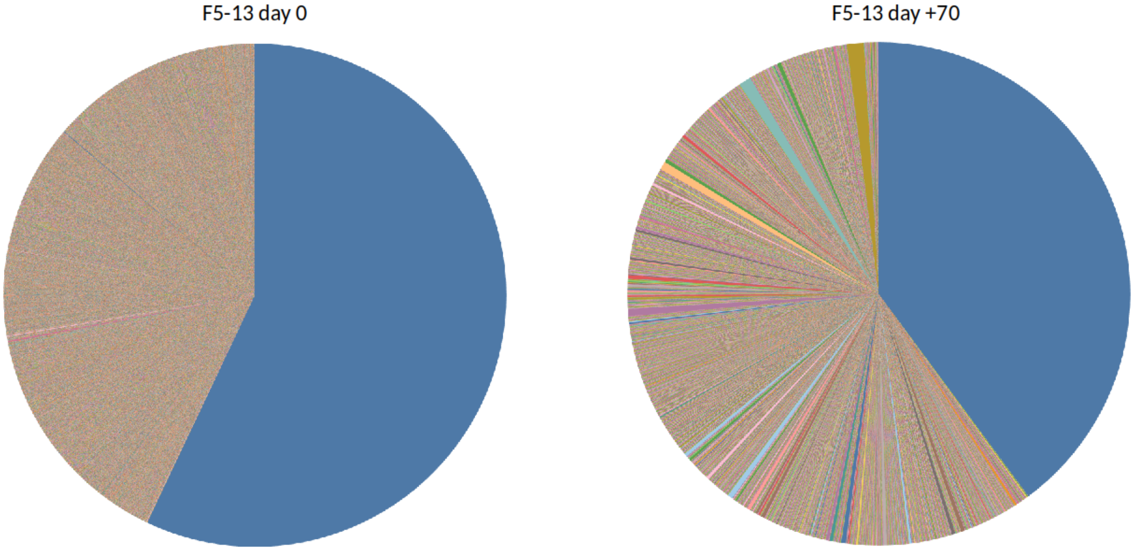


*

S.Figure1: Graphical representation of the MSCV retroviral vector which encodes the transgenic F5 MART-1 TCR and the NY-ESO-1 TCR used in our adoptive cell therapy clinical trials.

alignment scores (40, 80, 120, 160, 200, 240, 280). With a minimum alignment score of 160, 29.2% (STD 1.41%) of MART-1 TCR and 36.6% (STD. 2.17%) NY-ESO-1 alignments were called correctly as vector spanning reads with correct mapping coordinates, and 0.183% (STD 0.0947%) and 0.301% (STD 0.0141%) alignments were called incorrectly. At minimum, a minimum alignment score of 160 at least 80 read bases must map for a valid alignment. Simulated split reads with fewer than this threshold are undetectable, resulting in a large proportion of simulated reads being unobserved. However, when a vector spanning is called the vast majority are called correctly. The complete simulation and validation pipeline can be found at https://github.com/NuttyLogic/Epigenetic_Suppression_of_Transgenic_T-cell_Nowicki.2020/tree/master/VectorInsertionValidation.

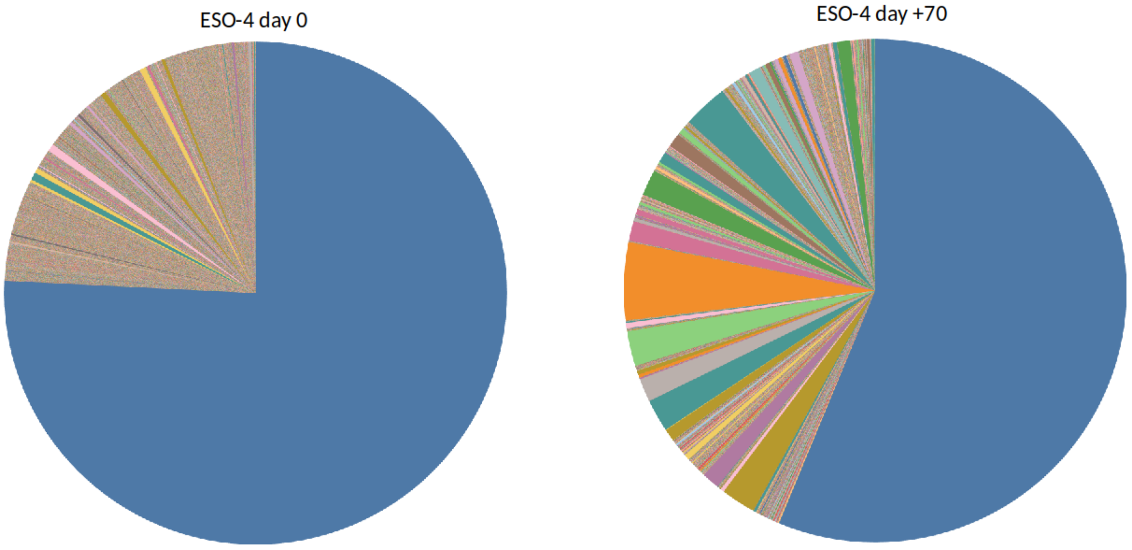
Figure S2.



*

S.Figure2: Representative TCR sequencing plots showing the relative proportion of transgenic F5-MART-1 TCR (Figure S2) and NY-ESO-1 TCR (Figure S3) at day 0 (infusion product) and day +70. T he transgenic TCR is indicated in blue, and each pie chart slice represents another individual TCR clonotype.

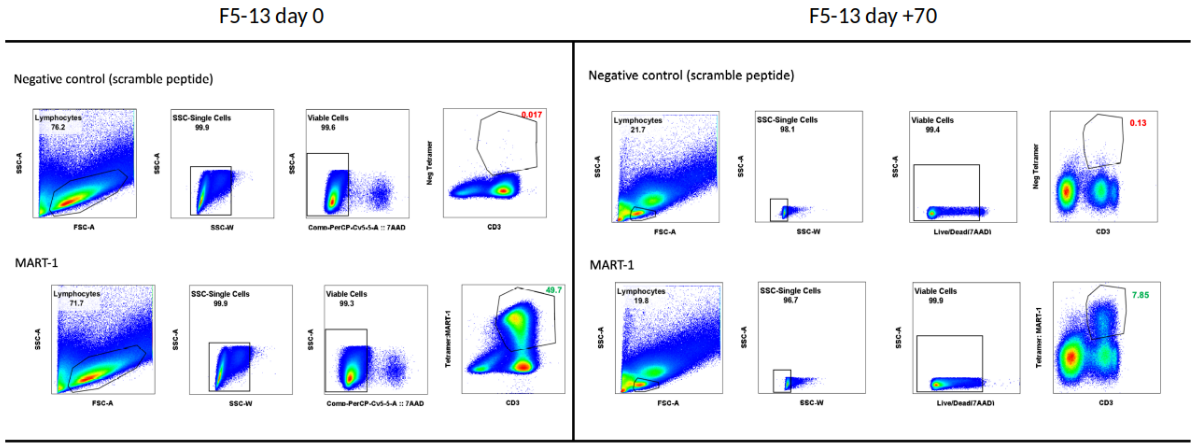
Figure S3.



*

S.Figure3: Representative TCR sequencing plots showing the relative proportion of transgenic F5-MART-1 TCR (Figure S2) and NY-ESO-1 TCR (Figure S3) at day 0 (infusion product) and day +70. The transgenic TCR is indicated in blue, and each pie chart slice represents another individual TCR clonotype.

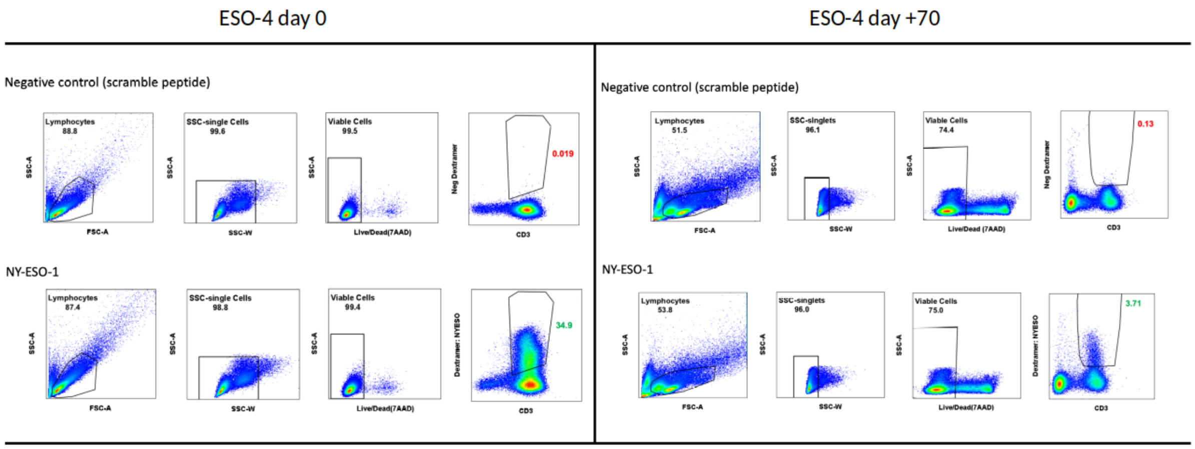
Figure S4.



*

S.Figure4: Representative fluorescent dextramer plots and gating strategies for the MART-1 (Figure S4) and NY-ESO-1 (Figure S5) TCRs at day 0 (infusion product) and day +70, and negative controls (scramble peptide).

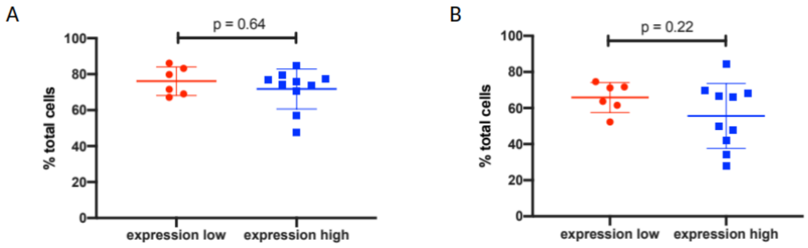
Figure S5.



*

S.Figure5: Representative fluorescent dextramer plots and gating strategies for the MART-1 (Figure S4) and NY-ESO-1 (Figure S5) TCRs at day 0 (infusion product) and day +70, and negative controls (scramble peptide).

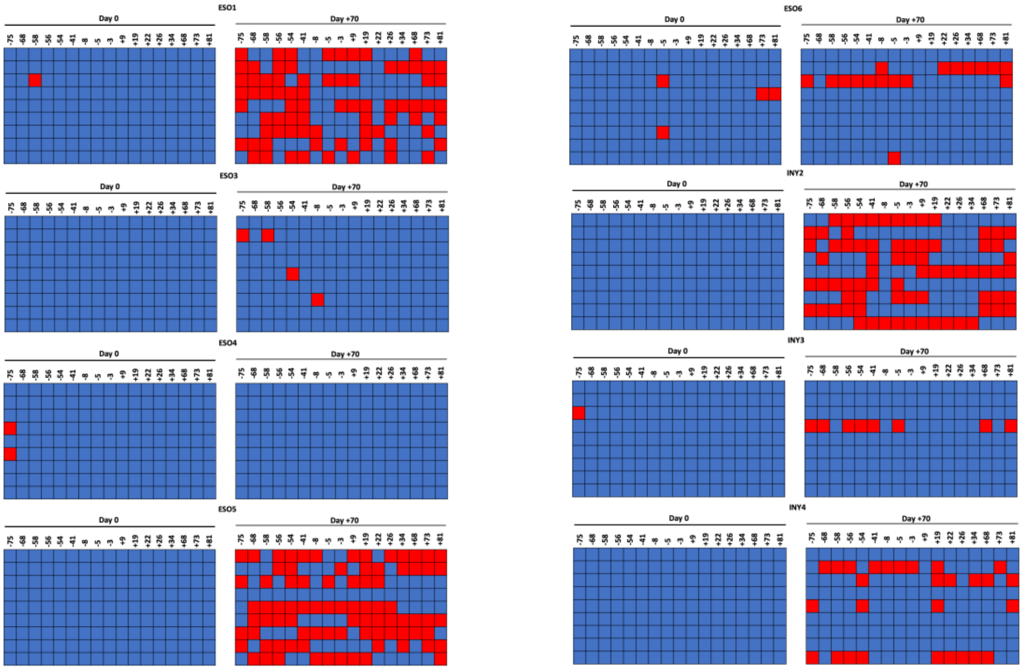
Figure S6



*

S.Figure6: Statistical comparisons between expression-high and expression-low patients' transgenic TCR DNA (A) and surface protein expression (B) as percentage of total cells at day 0 (baseline infusion product). Non-significant p-values are inset within each comparison (Mann-Whitney U test).

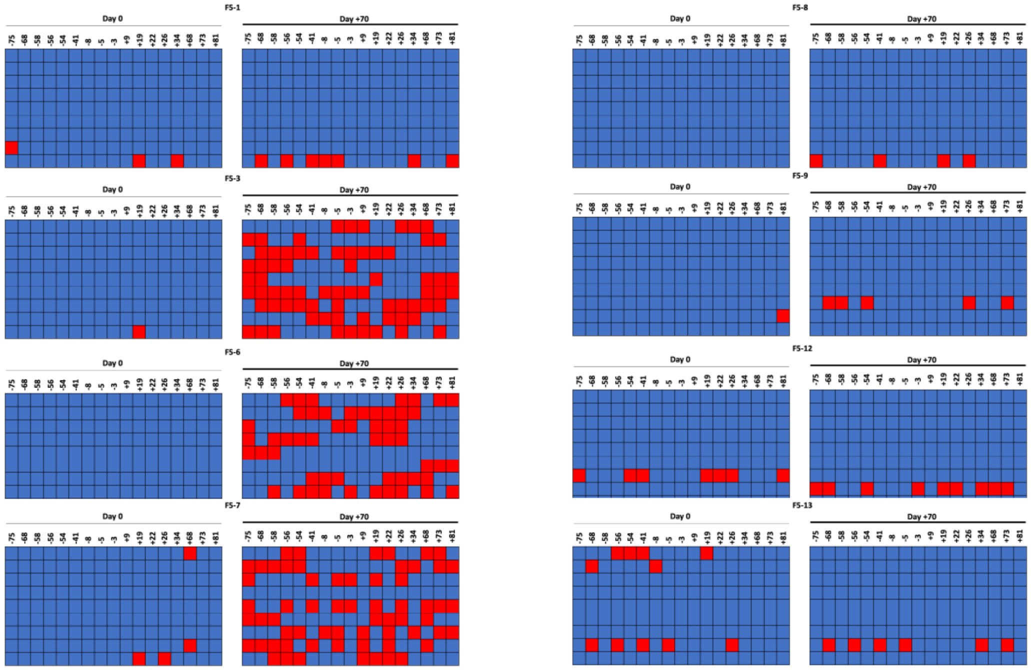
Figure S7



*

S.Figure7: Targeted bisulfite sequencing data of the CpG island within the MSCV 5'LTR promoter from all patients studied at day 0 (infusion product) and day +70. Patients treated with NY-ESO-1 TCR products are detailed in Figure S3, while patients treated with F-5 MART-1 TCR products are summarized in Figure S4. Each row represents sequencing of an individual experiment, with methylated cytosine loci indicated by red boxes and unmethylated loci by blue boxes. CpG loci positions are listed at the top of the graph relative to the TSS for the transgenic TCR.

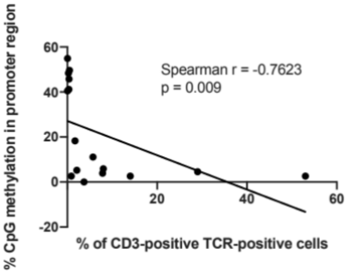
Figure S8



*

S.Figure8: Targeted bisulfite sequencing data of the CpG island within the MSCV 5'LTR promoter from all patients studied at day 0 (infusion product) and day +70. Patients treated with NY-ESO-1 TCR products are detailed in Figure S3, while patients treated with F-5 MART-1 TCR products are summarized in Figure S4. Each row represents sequencing of an individual experiment, with methylated cytosine loci indicated by red boxes and unmethylated loci by blue boxes. CpG loci positions are listed at the top of the graph relative to the TSS for the transgenic TCR.

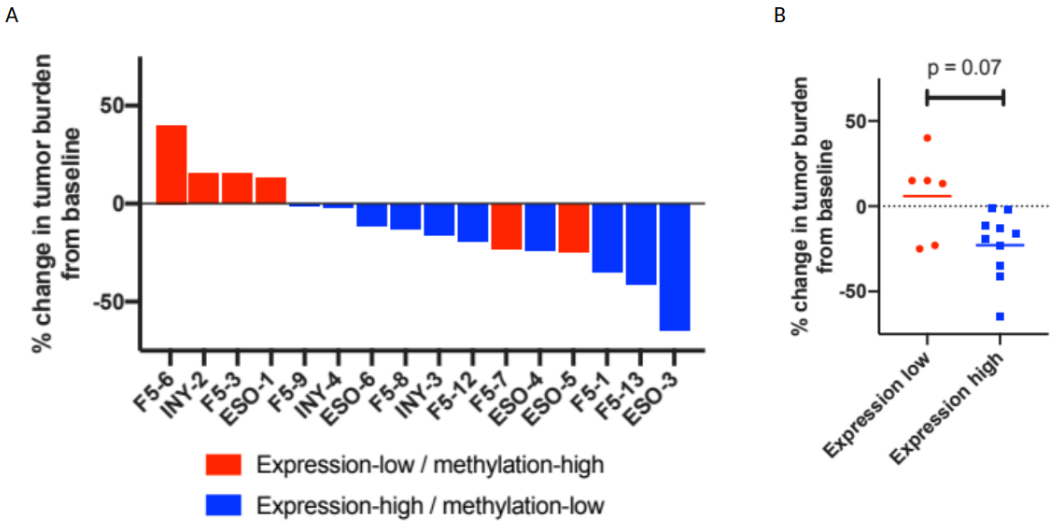
Figure S9



*

S.Figure9: Spearman correlation between percentage of CpG loci methylation within retroviral promoter region and surface expression of transgenic TCRs in day +70 samples for all patients.

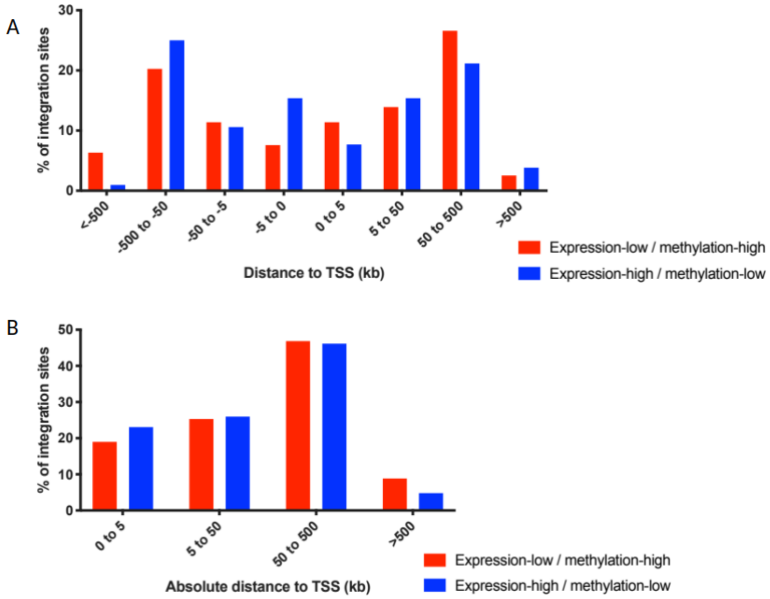
Figure S10



*

S.Figure10: A) Waterfall plot demonstrating percent change in tumor burden in each patient at end of study period. B) Comparison of change in tumor burden between expression-low and expression-high patient cohorts (Mann-Whitney U test).

Figure S11



*

S.Figure11: MSCV-TCR vector integration occurs sporadically throughout the host genome relative to transcription start sites. A) Proportion of MSCV-TCR vector integration sites sorted by distance to gene transcription start sites (TSS) compared between patients with high vs. low MSCV methylation values; data are sorted by absolute distance to TSS in B.

Bibliography

[1] Mari Aker et al. “Integration bias of gammaretrovirus vectors following transduction and growth of primary mouse hematopoietic progenitor cells with and without selection”. en. In: *Mol. Ther.* 14.2 (Aug. 2006), pp. 226–235.

[2] William R Burns et al. “Lack of specific γ-retroviral vector long terminal repeat promoter silencing in patients receiving genetically engineered lymphocytes and activation upon lymphocyte restimulation”. In: *Blood, The Journal of the American Society of Hematology* 114.14 (2009), pp. 2888–2899.

[3] Thinle Chodon et al. “Adoptive transfer of MART-1 T-cell receptor transgenic lymphocytes and dendritic cell vaccination in patients with metastatic melanoma”. en. In: *Clin. Cancer Res.* 20.9 (May 2014), pp. 2457–2465.

[4] Jeffrey Chou et al. “Epigenetic modulation to enable antigen-specific T-cell therapy of colorectal cancer”. en. In: *J. Immunother.* 35.2 (Feb. 2012), pp. 131–141.

[5] Begoña Comin-Anduix et al. *Definition of an Immunologic Response Using the Major Histocompatibility Complex Tetramer and Enzyme-Linked Immunospot Assays.* 2006.

[6] Sandra P D’Angelo et al. “Antitumor Activity Associated with Prolonged Persistence of Adoptively Transferred NY-ESO-1 c259T Cells in Synovial Sarcoma”. en. In: *Cancer Discov.* 8.8 (Aug. 2018), pp. 944–957.

[7] Suk See De Ravin et al. “Enhancers are major targets for murine leukemia virus vector integration”. en. In: *J. Virol.* 88.8 (Apr. 2014), pp. 4504–4513.

[8] Mark E Dudley et al. “Cancer regression and autoimmunity in patients after clonal repopulation with antitumor lymphocytes”. en. In: *Science* 298.5594 (Oct. 2002), pp. 850–854.

[9] Colin Farrell et al. “BiSulfite Bolt: A bisulfite sequencing analysis platform”. en. In: *Gigascience* 10.5 (May 2021).

[10] R G Hawley et al. “Versatile retroviral vectors for potential use in gene therapy”. en. In: *Gene Ther.* 1.2 (Mar. 1994), pp. 136–138.

[11] Naomi N Hunder et al. “Treatment of metastatic melanoma with autologous CD4+ T cells against NY-ESO-1”. en. In: *N. Engl. J. Med.* 358.25 (June 2008), pp. 2698–2703.

[12] Carl H June and Michel Sadelain. *Chimeric Antigen Receptor Therapy.* 2018.

[13] P L Klarenbeek et al. “Inflamed target tissue provides a specific niche for highly expanded T-cell clones in early human autoimmune disease”. en. In: *Ann. Rheum. Dis.* 71.6 (June 2012), pp. 1088–1093.

[14] Yuichi Kumaki, Masaaki Oda, and Masaki Okano. “QUMA: quantification tool for methylation analysis”. en. In: *Nucleic Acids Res.* 36.Web Server issue (July 2008), W170–5.

[15] Heng Li et al. “The Sequence Alignment/Map format and SAMtools”. en. In: *Bioinformatics* 25.16 (Aug. 2009), pp. 2078–2079.

[16] Long-Cheng Li and Rajvir Dahiya. “MethPrimer: designing primers for methylation PCRs”. en. In: *Bioinformatics* 18.11 (Nov. 2002), pp. 1427–1431.

[17] Richard A Morgan et al. “Cancer regression in patients after transfer of genetically engineered lymphocytes”. en. In: *Science* 314.5796 (Oct. 2006), pp. 126–129.

[18] Theodore S Nowicki et al. “A Pilot Trial of the Combination of Transgenic NY-ESO-1–reactive Adoptive Cellular Therapy with Dendritic Cell Vaccination with or without Ipilimumab”. en. In: *Clin. Cancer Res.* 25.7 (Apr. 2019), pp. 2096–2108.

[19] Alexander Pfeifer et al. “Transgenesis by lentiviral vectors: lack of gene silencing in mammalian embryonic stem cells and preimplantation embryos”. en. In: *Proc. Natl. Acad. Sci. U. S. A.* 99.4 (Feb. 2002), pp. 2140–2145.

[20] Seth M Pollack et al. “NYESO-1/LAGE-1s and PRAME are targets for antigen specific T cells in chondrosarcoma following treatment with 5-Aza-2-deoxycitabine”. en. In: *PLoS One* 7.2 (Feb. 2012), e32165.

[21] Paul F Robbins et al. “A Pilot Trial Using Lymphocytes Genetically Engineered with an NY-ESO-1–Reactive T-cell Receptor: Long-term Follow-up and Correlates with Response”. en. In: *Clin. Cancer Res.* 21.5 (Mar. 2015), pp. 1019–1027.

[22] Theodore L Roth et al. *Reprogramming human T cell function and specificity with non-viral genome targeting.* 2018.

[23] C Scott Swindle, Hyung G Kim, and Christopher A Klug. “Mutation of CpGs in the murine stem cell virus retroviral vector long terminal repeat represses silencing in embryonic stem cells”. en. In: *J. Biol. Chem.* 279.1 (Jan. 2004), pp. 34–41.

[24] Dimitrios N Vatakis et al. “Antitumor activity from antigen-specific CD8 T cells generated in vivo from genetically engineered human hematopoietic stem cells”. en. In: *Proc. Natl. Acad. Sci. U. S. A.* 108.51 (Dec. 2011), E1408–16.

[25] James C Yang and Steven A Rosenberg. “Adoptive T-Cell Therapy for Cancer”. en. In: *Adv. Immunol.* 130 (Feb. 2016), pp. 279–294.

[26] Lili Yang and David Baltimore. “Long-term in vivo provision of antigen-specific T cell immunity by programming hematopoietic stem cells”. en. In: *Proc. Natl. Acad. Sci. U. S. A.* 102.12 (Mar. 2005), pp. 4518–4523.

[27] Shuyuan Yao et al. “Retrovirus silencing, variegation, extinction, and memory are controlled by a dynamic interplay of multiple epigenetic modifications”. en. In: *Mol. Ther.* 10.1 (July 2004), pp. 27–36.

CHAPTER 4

The Epigenetic Pacemaker - modeling epigenetic states under an evolutionary framework.

Colin Farrell[1], Sagi Snir[2,4], Matteo Pellegrini[3,4]

[1]Department of Human Genetics, University of California, Los Angeles, CA, USA;

[2]Dept. of Evolutionary Biology, University of Haifa, Israel;

[3]Dept. of Molecular, Cell and Developmental Biology; University of California, Los Angeles, CA 90095, USA;

[4]Corresponding Authors, These authors contributed equally; matteop@mcdb.ucal.edu, ssagi@research.haifa.ac.il

Epigenetic rates of change, much as evolutionary mutation rate along a lineage, vary during lifetime. Accurate estimation of the epigenetic state has vast medical and biological implications. To account for these nonlinear epigenetic changes with age, we recently developed a formalism inspired by the Pacemaker model of evolution that accounts for varying rates of mutations with time. Here, we present a python implementation of the Epigenetic Pacemaker (EPM), a conditional expectation maximization algorithm that estimates epigenetic landscapes and the state of individuals and may be used to study nonlinear epigenetic aging. The EPM is available at https://pypi.org/project/EpigeneticPacemaker/.

4.1 Introduction

Methylation of cytosine plays an integral role in the regulation of gene expression and mammalian development[7, 10][9, 18]. During the mammalian life cycle, age associated changes in DNA methylation proceed predictably and nonlinearly with time[14]. The systematic changes of DNA methylation with age have led to the development of several epigenetic clocks[4, 3]. Most of these models assume that the change in methylation is linear with age, and as such are reminiscent of the molecular clock concept in molecular evolution. The predicted age from these models can be interpreted as a physiological or epigenetic age, and the residual error between the expected and predicted epigenetic age has been associated with several health outcomes[5, 6, 12]. However, these approaches make a priori assumptions about the functional relationship between epigenetic changes and age (e.g linearity) and hence may fail to adequately capture nonlinear changes in methylation with age. This is important because there is substantial evidence to suggest that epigenetic changes are much more rapid early in life and progressively slow as we age across tissue type [14].

To overcome the limitations of prior approaches we developed an evolutionary based approach - the Epigenetic Pacemaker (EPM) - for modeling epigenetic states as evolving entities[16, 15]. The EPM borrows from the Universal Pacemaker formalism (UPM)[17] under which the evolutionary rate of genes remains constant relative to one another but the absolute rate can change arbitrarily by factors affecting the evolving lineage. In contrast to the EPM, most previous epigenetic clocks resemble the molecular evolutionary concept of the Molecular Clock[21], where the evolutionary rate of genes remains constant with time. In the EPM, given a set of i methylation sites and j individuals, the observed methylation status, $\hat{m}_{ij}$, is given as $\hat{m}_{ij} = m_i^0 + r_i s_j + \epsilon_{ij}$, where m_i^0 is the initial methylation value, r_i is the rate of methylation change, s_j is the epigenetic state, and ϵ_{ij} is a normally distributed error term. Given an input matrix $\hat{M} = [\hat{m}_{ij}]$ the goal of the EPM is to find the optimal values of r_i, m_i^0, and s_j to minimize the error between the measured and predicted methylation values. Our approach is distinct from previous epigenetic clock methods that attempt to minimize the difference between observed and predicted age, thus implicitly constraining the functional form of that relationship.

The EPM optimization is accomplished through an implementation of a fast conditional expectation maximization algorithm that we have previously shown maximizes the model likelihood by minimizing the residual sum of squares error(Snir et al. 2016). When fitting the EPM each methylation site is assigned an independent rate of change, and starting methylation value, and each individual is assigned an epigenetic state. We use chronological age as an initial guess for the epigenetic state, which is then updated through each iteration to minimize the error across the observed epigenetic landscape (i.e. the parameter set of our model). Because we model methylation and not age, the EPM relaxes the condition of linearity between a trait of interest (e.g. age) and the observed methylation values. This allows the EPM to model non-linear relationships between our state, s_j, and the trait, without needing to transform the trait of interest (as is done in certain epigenetic clocks).

4.2 Epigenetic Pacemaker

To highlight the utility of the EPM, we fit EPM and linear regression models using publicly available Illumina HumanMethylation450 (450k) microarray data [13] generated from human brain tissue samples ($n = 675, 0 - 96 years$) [7]. Briefly, we performed stratified sampling by age to select 270 brain tissue samples for site selection and model training. CpG sites were selected for model inclusion using the absolute value of the Pearson correlation coefficient between the training methylation values and chronological age, ($PCC \geq |0.85|, n = 254$). We then fit the EPM and regression models[13, 2, 11] using the selected sites and training methylation data. Age and epigenetic state predictions were made for the remaining brain tissue samples ($n = 405$) left out of model training.The EPM model shows the non-linear relationship between epigenetic state and chronological age (Figure 1A) that is lost in the regression model (Figure 1B). We then used the brain EPM and linear models to predict the epigenetic state of 450k data ($n = 732, 14 - 96 years$) generated from whole blood tissue [8]. Samples with missing methylation values for the CpG sites used in model generation were dropped, resulting in 634 analysis samples. The brain EPM model captures aging in the whole blood samples with minimal error (Figure 1C), while the aging signal is largely lost in the linear model (Figure 1D).

We have developed an optimized version of the EPM algorithm implemented as a python package[20][19][1] that adopts Scikit-Learn[11] style syntax for easy incorporation into current workflows with support for cross validation. The EPM is available through the python package repository, https://pypi.org/project/EpigeneticPacemaker/, under a MIT license. Full documentation, including tutorials, and source code can be found at https://epigeneticpacemaker.readt and https://github.com/
NuttyLogic/EpigeneticPacemaker respectively

4.3 Availability and requirements

Project name: Epigenetic Pacemaker
Project home page: https://github.com/NuttyLogic/EpigeneticPacemaker
Operating system(s): Platform independent
Programming language: Python ¿= 3.6
Other requirements: numpy¿=1.16.3, tqdm¿=4.31.1, scipy¿=1.3.0
License: MIT

Acknowledgments and Funding

This work was supported by the National Institutes of Health (T32CA201160 to C.F.).

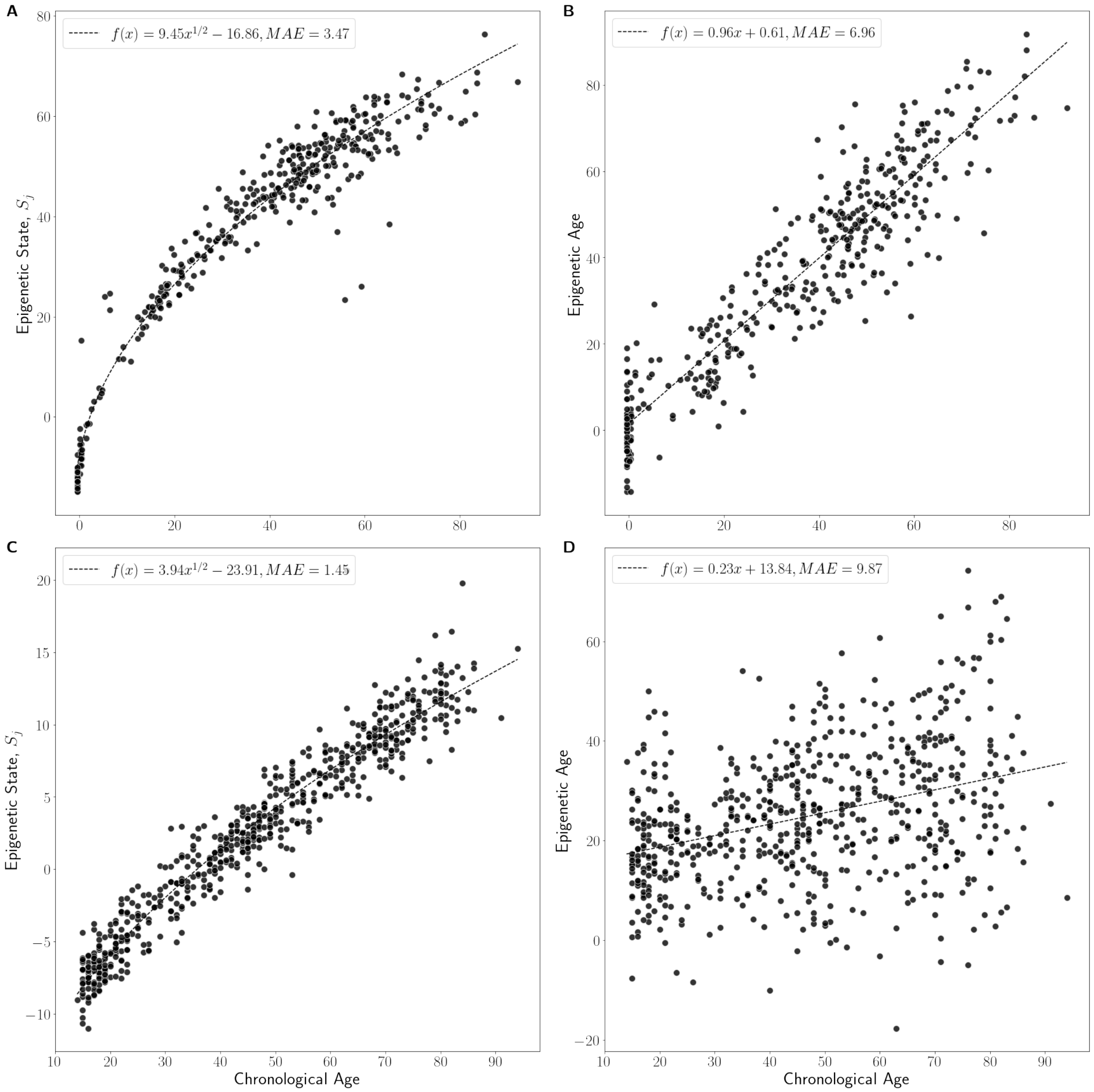


Figure1: Epigenetic state predictions for $(n = 405)$ test samples compared to the chronological age of each sample with a line of best fit for the EPM (A) and linear regression (B) models. The non-linear trend observed in the EPM model better captures the observed aging trend and reduces observed error as measured by mean absolute error (MAE). (C) Epigenetic state predictions made for whole blood samples using the EPM and (D) linear model.

Bibliography

[1] Casper O da Costa-Luis. “tqdm: A Fast, Extensible Progress Meter for Python and CLI”. In: *JOSS* 4.37 (May 2019), p. 1277.

[2] Raul Garreta and Guillermo Moncecchi. *Learning scikit-learn: Machine Learning in Python*. en. Packt Publishing Ltd, Nov. 2013.

[3] Gregory Hannum et al. “Genome-wide methylation profiles reveal quantitative views of human aging rates”. en. In: *Mol. Cell* 49.2 (Jan. 2013), pp. 359–367.

[4] Steve Horvath. “DNA methylation age of human tissues and cell types”. en. In: *Genome Biol.* 14.10 (2013), R115.

[5] Steve Horvath and Andrew J Levine. “HIV-1 Infection Accelerates Age According to the Epigenetic Clock”. en. In: *J. Infect. Dis.* 212.10 (Nov. 2015), pp. 1563–1573.

[6] Steve Horvath et al. “Accelerated epigenetic aging in Down syndrome”. en. In: *Aging Cell* 14.3 (June 2015), pp. 491–495.

[7] Andrew E Jaffe et al. “Mapping DNA methylation across development, genotype and schizophrenia in the human frontal cortex”. en. In: *Nat. Neurosci.* 19.1 (Jan. 2016), pp. 40–47.

[8] Asa Johansson, Stefan Enroth, and Ulf Gyllensten. “Continuous Aging of the Human DNA Methylome Throughout the Human Lifespan”. en. In: *PLoS One* 8.6 (June 2013), e67378.

[9] E Li, T H Bestor, and R Jaenisch. “Targeted mutation of the DNA methyltransferase gene results in embryonic lethality”. en. In: *Cell* 69.6 (June 1992), pp. 915–926.

[10] Masaki Okano et al. *DNA Methyltransferases Dnmt3a and Dnmt3b Are Essential for De Novo Methylation and Mammalian Development.* 1999.

[11] F Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *J. Mach. Learn. Res.* 12 (2011), pp. 2825–2830.

[12] Laura Perna et al. *Epigenetic age acceleration predicts cancer, cardiovascular, and all-cause mortality in a German case cohort.* 2016.

[13] Juan Sandoval et al. "Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome". en. In: *Epigenetics* 6.6 (June 2011), pp. 692–702.

[14] Sagi Snir, Colin Farrell, and Matteo Pellegrini. "Human epigenetic ageing is logarithmic with time across the entire lifespan". en. In: *Epigenetics* 14.9 (Sept. 2019), pp. 912–926.

[15] Sagi Snir and Matteo Pellegrini. "An epigenetic pacemaker is detected via a fast conditional expectation maximization algorithm". en. In: *Epigenomics* 10.6 (June 2018), pp. 695–706.

[16] Sagi Snir, Bridgett M vonHoldt, and Matteo Pellegrini. "A Statistical Framework to Identify Deviation from Time Linearity in Epigenetic Aging". en. In: *PLoS Comput. Biol.* 12.11 (Nov. 2016), e1005183.

[17] Sagi Snir, Yuri I Wolf, and Eugene V Koonin. *Universal Pacemaker of Genome Evolution.* 2012.

[18] Peri H Tate and Adrian P Bird. *Effects of DNA methylation on DNA-binding proteins and gene expression.* 1993.

[19] Pauli Virtanen et al. "SciPy 1.0: fundamental algorithms for scientific computing in Python". In: *Nat. Methods* (Feb. 2020).

[20] Stéfan van der Walt et al. *The NumPy Array: A Structure for Efficient Numerical Computation.* 2011.

[21] Emile Zuckerkandl and Linus Pauling. *Evolutionary Divergence and Convergence in Proteins.* 1965.

CHAPTER 5

Human Epigenetic Aging is Logarithmic with Time across the Entire LifeSpan

Sagi Snir[1], Colin Farrell[2], and Matteo Pellegrini[3,4]

[1]Dept. of Evolutionary Biology, University of Haifa, Israel;

[2] Department of Human Genetics, University of California Los Angeles, Los Angeles, CA 90095, USA.

[3]Dept. of Molecular, Cell and Developmental Biology; University of California, Los Angeles, CA 90095, USA;

[4]Corresponding Author; matteop@mcdb.ucla.edu

Epigenetic changes during aging have been characterized by multiple epigenetic clocks, that allow the prediction of chronological age based on methylation status. Despite their accuracy and utility, epigenetic age biomarkers leave many questions about epigenetic aging unanswered. Specifically, they do not permit the unbiased characterization of non-linear epigenetic aging trends across entire life spans, a critical question underlying this field of research. Here we a provide an integrated framework to address this question. Our model, inspired from evolutionary models, is able to account for acceleration/deceleration in epigenetic changes by fitting an individual's model age, the epigenetic age, which is related to chronological age in a non-linear fashion. Application of this model to DNA methylation data measured across broad age ranges, from before birth to old age, and from two tissue types, suggests a universal logarithmic trend characterizes epigenetic aging across entire lifespans.

5.1 Introduction

Cell type specific differences in gene expression are partially controlled by chromatin accessibility and specific covalent modifications. These include modification to histones and DNA [35, 4]. Among these, DNA methylation has been one of the most extensively studied components of cell type specification [16, 8]. The covalent attachment of a methyl group to cytosine is catalyzed by either de novo or maintenance methyltransferases, and in mammals is primarily targeted to CpG dinucleotides. Most CpGs in mammalian genomes are methylated, but pockets of hypomethylation exist, largely at promoters and enhancers. It has been shown that the absence of DNA methylation is closely associated with the presence of H3K4 methylation, which is also a hallmark of enhancers and promoters [23]. As stem cells differentiate along myriad lineages, each cell type tends to have distinctive DNA methylation profiles, mostly due to differential activation of enhancers [12].

Once an organism reaches its adult stage, these cell types and their respective epigenomes, have been largely determined. While these developmental epigenetic changes are believed to be rapid and extensive, in the past few years it has become ever more apparent that epigenetic changes continue to occur as an organism ages [29]. This observation has led to the development of multiple epigenetic clocks, that is, biomarkers that accurately predict the chronological age of an animal based on its DNA methylation profile [13, 11]. These epigenetic clocks have been extensively used in aging research and have proven to be more accurate than previous aging biomarkers, such as the length of telomeres [20]. Using these epigenetic clocks, much has been learned about the effects of the environment on aging. For example, it is well known that the restriction of calories in mice slows down aging, increases lifespan as well as the rate of the epigenetic clock [32]. Similar conclusions have been found in humans, where individuals with more rapid epigenetic aging tend to suffer from higher all-cause mortality rates [1].

While these epigenetic clocks have proven to be useful for aging research, they are constructed using machine learning methods that provide limited insights into the underlying processes that are driving these changes. For example, the Horvath epigenetic clock biomarker is constructed by selecting 354 CpG sites using penalized lasso regression, that optimally predict the chronological age of an individual. This biomarker also sets a hardcoded boundary of 20 years, where childhood ages are transformed using a logarithmic function up to this boundary, while adult aging (above the 20 boundary) is kept linear [13]. This biomarker generates very accurate predictions of chronological age, typically within a couple of years, but leaves many questions unanswered. Is there truly a change in the rate of epigenetic aging, from logarithm to linear trends, at 20 years? Does the linear fit of epigenetic age persist indefinitely, even for older individuals? Do non-linear trends in epigenetic aging vary across populations? As the biomarkers are species specific, they also do not allow one to directly address whether epigenetic aging trends vary across species.

To address some of these questions, we have previously proposed a common framework by borrowing from the field of evolution. The universal pacemaker (UPM) of genome evolution was devised in the setting of molecular evolution in order to relax the time-linear evolution (i.e. rate constancy) imposed by the molecular clock hypothesis [25, 34, 24], to account for correlation between rate changes in the genes of an evolving organism. The UPM is a statistical framework under which the relative evolutionary rates of all genes remain nearly constant whereas the absolute rates can change arbitrarily. In [27] we first proposed the adaptation of the UPM to the epigenetic setting, named the *epigenetic pacemaker* (EPM), and showed its application to simulation and small scale biological data. To the best of our knowledge, the EPM is the first model based framework for epigenetic aging, where the rates of change with time of individual CpG sites are parametrized, along with the epigenetic age of the individual. In [26] we devised a fast, conditional expectation maximization (CEM) algorithm that is capable of processing inputs of several thousands of sites and individuals.

In this work we set out to address some of the questions mentioned above, regarding the non-linear trends of epigenetic aging across populations. First we show that the EPM, which lacks any predefined regimes of age intervals, can be used to model and identify epigenetic aging trends over the entire lifespan of a population. We first apply our approach to a synthetic model simulating a non-linear aging process and show that our framework is capable of capturing the trend built into this model. Next we apply our model to publicly available sets of DNA methylation collected across broad age ranges and diverse tissues. Our results suggest unambiguously that a logarithmic trend across the entire lifespan is a better description of epigenetic aging than linear or polynomial trends.

5.2 Methods

The Evolutionary Models

Our basic objects are a set of m *individuals* and n *methylation sites* in a genome (or simply sites). Each individual has an age, forming the set t of *time periods* $\{t_j\}$ corresponding to each individual j's age. Henceforth we will interchangeably refer to individuals with their age. Each individual has a set of sites s_i undergoing methylation changes at some *characteristic rate* r_i. Each site s_i starts at some *methylation start level* s_i^0. All individuals have all the sites s_i. As r_i and s_i^0 are characteristic of the site s_i, by the model they are the same at all individuals. The latter fact, links the same sites across different individuals, but also within individuals by the fact that sites generally maintain the same characteristic rates across the whole population. Henceforth, we will index sites with i and individuals with j.

Now, let $s_{i,j}$ measure the methylation level at site s_i in individual j after time (i.e. age) t_j. Hence, under the *molecular clock* model (MC), where rate of change is relatively constant over time, we expect: $s_{ij} = s_i^0 + r_i t_j$. However, in reality we have a *noise effect* $\varepsilon_{i,j}$ that is added and therefore the *observed value* $\hat{s}_{ij}$ is $\hat{s}_{ij} = s_i^0 + r_i t_j + \varepsilon_{i,j}$.

Our goal is to find, given the input matrix $\hat{S} = [\hat{s}_{i,j}]$, the maximum likelihood (ML) values for the variables r_i and s_i^0 for $1 \leq i \leq n$. For this purpose, we assume a statistical model for $\varepsilon_{i,j}$ by assuming that it is normally distributed, $\varepsilon_{i,j} \sim N(0, \sigma^2)$. In [27] we showed that minimizing the following function, denoted RSS, is equivalent to maximizing the model's likelihood

$$RSS = \sum_{i \leq n} \sum_{j \leq m} (\hat{s}_{i,j} - (s_i^0 + r_i t_j))^2. \tag{5.1}$$

We also showed that there is an efficient and precise linear algebra solution to this problem, that we describe in more detail in the supplementary text.

In contrast to the MC, under the EPM model, sites may arbitrarily and independently of their counterparts in other individuals, change their rate at any point in life. However, when this happens, all sites of that individual change their rate proportionally such that the ratio $r_i/r_{i'}$ is constant between any two sites i, i' at any individual j and at all times. In [27] we showed that this is equivalent to extending individual j's age by the same proportion of the rate change. The new age is denoted as the *epigenetic age*. Therefore here we do not just use the given chronological age but estimate the age of each individual. Hence under the EPM we must find the optimal values of s_i^0, r_i, and t_j (where t_j represents a weighted average of the rate changes an individual has undergone through life). The solution to this optimization problem is described in detail in our previous publications [27, 26]. We note that the deviation between the chronological age and the estimated epigenetic age is an age difference which, when positive, is denoted as age acceleration, and deceleration - otherwise.

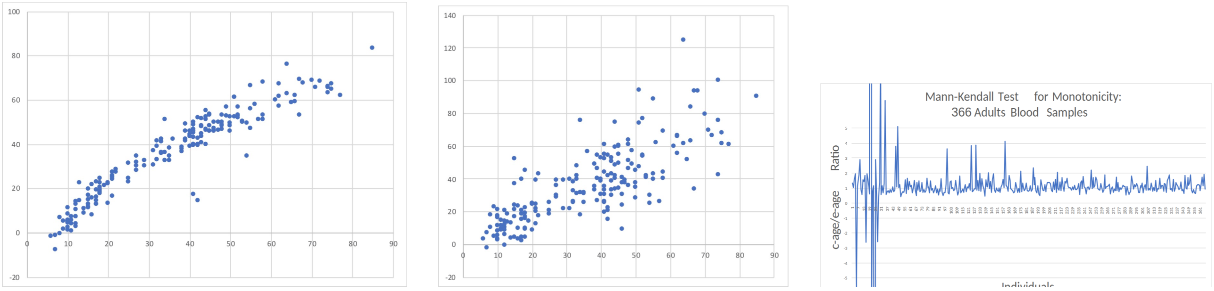
To compare between the two models - MC and EPM, we note the following. The MC model is restricted to linearity with time by estimating a constant rate of methylation at each site, and using the given chronological age of each individual. The competing, relaxed, model (EPM) has no such restriction, and we estimate an ”epigenetic” age for each individual. By definition, the ML solution under the relaxed model cannot be worse than the constrained model. For that specific case, when one hypothesis generalizes another, there is a special test, the likelihood ratio test (LRT), in which the specific hypothesis serves as the null hypothesis and the goal is to reject it in favor of the alternative one. In the supplementary text we provide a more detailed explanation of this test and its application in our case.

Selecting Informative Methylation Loci

DNA methylation platforms usually measure several hundreds of thousands of sites. It has been observed that many of these sites are invariant and do not change with age. It is desirable to restrict the analysis only to the most informative sites. Nevertheless, among the sites that do change it is necessary to set a criterion for site selection as it is inefficient to analyze all of the sites. There are several alternatives that we now describe.

The first and most basic and intuitive criterion for site selection is site variance - simply choose the sites that exhibit the largest variability. Figure 1(L) depicts the resulted analysis based on this criterion. This criterion is crude in the sense that it entirely ignores the relationship between time (age) and methylation. Therefore the next criterion to be examined is the covariance between age and methylation status at the site. The covariance metric selects sites that have a large change in methylation with age. We note that this criterion will not necessarily yield a significant linear fit between age and methylation status, the sites may still have a significant scatter, as is shown in Figure 1(M).

Therefore the third criterion is the (absolute) Pearson correlation coefficient (PCC) defined as $\rho_{X,Y} = Cov(X,Y)/\sigma_X/\sigma_y$. In contrast to the covariance, PCC selects sites that have a tight fit to a linear relationship between methylation and age, although some of them show small changes in methylation across the range. Figure 1 (R) shows the results based on sites selected by the PCC criterion. It is noticeable that using this criterion a much tighter relationship between epigenetic and chronological age is obtained. Hence we use PCC to select the sites to be modeled for all our data sets as it provided the clearest trends between epigenetic and chronological age.



(a) - Max Variance

(b) - Max Covariance

(c) Max PCC

Figure 1: Site Selection Criterion

Scatter plots of inferred epigenetic age (e-age, y-axis) as a function of the chronological age (c-age, x-axis) as a result of applying the EPM algorithm to blood samples from data set GSE60132 (see more details in the Results sec.). Each point represents an individual. 1000 best sites were selected by the following three criteria. (A) Sites are selected based on their variance, regardless of correlation to age. (B) Sites are selected based on their covariance with age. (C) Sites are selected by the (absolute) Pearson correlation coefficient.

Determining the trend line of epigenetic age

To determine the trend line between epigenetic and chronological age we employed both coarse and fine grained procedures based on the individual e-age inferred initially by the pacemaker criterion. Recall though that a first stage test is whether the pacemaker criterion, stating that rates and starting states are statistically correlated across individuals, and sites at any individual are also correlated, holds. This, is done by comparing to the molecular clock to the pacemaker model, as was described in the model description. Indeed, in the supplementary text we show the results of this test, along with the specific values obtained. The values depict that the pacemaker alternative is always superior with p-value smaller than 10^{-6}.

We start by describing the two-stage procedure for determining the type of trend in the population. The mutual independence of the stages, along with lack of any prior assumption of any trend in the population, guarantees that the trend inferred is objective and unbiased. The EPM procedure is applied to the data in order to find optimal values for rates, starting states, and epigenetic age for all sites and individuals. Note that except for that pacemaker principle, that enforces uniformity of site rate and starting state across all individuals, there is no mechanism imposing correlation between any two individuals. Therefore, the EPM assigns every individual the optimal epigenetic age.

Once the EPM procedure is done, each individual is assigned its own epigenetic age. At the second stage, we seek a function that best fits the relationship between epigenetic and chronological age across the entire age range. Our prime criterion for goodness of fit for a trend of this relationship between epigenetic and chronological age, is the R^2 coefficient from the trend line. In all our data sets we parametrized three functional forms for the trend line: linear, quadratic and exponential, and we used Excel to fit the best coefficients for each type. In the Results section we provide a more detailed description on this process.

We now describe a second approach that we devised and utilized. In age ranges where the trend is not conspicuous, that is, near linear, such that it cannot be distinguished convincingly using the above functional forms, we note the following. For a person j with age t_j and inferred e-age p_j we define $\rho_j = t_j/p_j$. Now, assume the increase in e-age is decreasing with time, then we observe that if we order the ρ_js by increasing t_j, we obtain a monotonic increasing series (of ρ_j). We denote that series $[\rho_j]$. However, due to biological and statistical noise we never expect to find strict monotonicity at $[\rho_j]$ and we are bound to test only a *trend* of monotonicity. Now, we note that by the definition of ρ_j also the variance of ρ_j is changing in time. However, we note the following. For any two indices j_1 and j_2 such that $j_1 < j_2$, if $[\rho_j]$ is monotonically decreasing, then $\rho_{j_1} < \rho_{j_2}$. Moreover, suppose we randomize the order of $[\rho_j]$. Then for any two indices j_1 and j_2 the probability $\P[\rho_{j_1} > \rho_{j_2}] = 1/2$ and hence the expected number of $j_1 < j_2$ such that $\rho_{j_1} < \rho_{j_2}$ is $\binom{n}{2}/2$. Now, since the variables ρ_j might be dependent, we cannot use standard bounds on deviations to calculate the probability of seeing that many pairs $j_1 < j_2$ such that $\rho_{j_1} < \rho_{j_2}$ by chance. This forces the use of a non parametric test of the hypothesis. For this purpose we can use the Mann-Kendall test for monotonic trend [**Kendall-1975**, 19, 10]. According to this method, given a random vector v, all pairs of indices (i, j) such that $i < j$ are checked whether $v_i < v_j$. Let p be the number of pairs (i, j) for $i < j$, such that $v_i < v_j$ and let q be the number of such pairs such that $v_i > v_j$. Now let $S = p - q$, representing first the direction of trend with $S > 0$ when the series is increasing, and vice versa for $S < 0$. However, S also indicates on the intensity of the trend, and we note that under h_0 (no monotonicity), we have $E(S) = 0$. Now we also need to compute the variance of S, $\mathrm{Var}(S) = \frac{n}{18}(n-1)(2n+5)$. The statistic z defined $z = \frac{S-1}{\sqrt{\mathrm{Var}(S)}}$ follows approximately the standard normal distribution, hence allowing us to obtain conveniently a p-value for the trend indicated by S.

5.3 Results

Identifying Trends in a Cohort

Methylation trends of the relationship between epigenetic and chronological time in a cohort provide useful information of how a group, as opposed to an individual, ages epigenetically with time. The Horvath model [13] has a rigid assumption of linearity of e-age in time for adults by using a linear combination of an individual's methylation states of several hundreds of sites. For kids (age less than 20), the model corrects for non linearity using a logarithmic, yet fixed, function. The EPM model has no such assumption and therefore has the freedom to assign each individual its own e-age, as long as it complies with the EPM universality law, that is, that this age affects all the individual's sites. We now demonstrate on synthetic data, the ability of our procedure to infer correct times (e-ages) and in particular trends throughout a whole population. For this purpose we have devised the following age related function that appears to encompass the characteristics of e-aging as they emerge from existing knowledge, in particular by the Horvath model:

$$p = c^{tr} * \frac{h}{h^{tr}} \tag{5.2}$$

where p holds the e-age, c is the chronological age (c-age) of a person, h is some upper limit on a person's age, and tr is a *trend* parameter to the function. The trend function has few desired characteristics. First, it satisfies a monotonic decrease in rate through time (c-age) and that decrease is proportional to the trend parameter. Also, at c-age that equals the upper limit h, the epigenetic and chronological ages coincide: $c = p$. Finally, for $tr = 1$ the trend function is linear with $p = c$ for every c, as $\frac{h}{h^{tr}} = 1$. Figure 2(L) illustrates pictorially the behavior of the trend function for several values of trend $tr = 1, 0.8, 0.5, 0.1$ and for upper limit age $h = 100$. Indeed we see that all trend lines depart from the origin and converge towards the point $c = h$. We also see that the larger tr is (with maximum $tr = 1$), the more straight the trend line is, and in particular, for $tr = 1$ a straight line with slope 1 is exhibited.

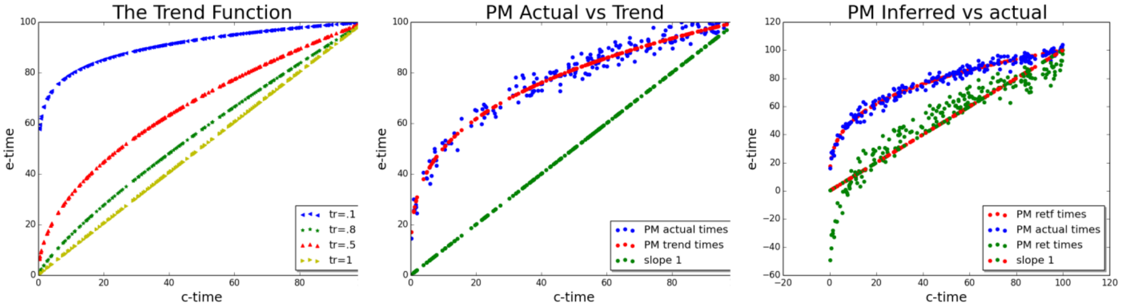


Figure 2 The Trend Function

(Left) Trend lines for four tr values $tr = 0.1, 0.5, 0.8, 1$ in blue, red, green, and olive green colors respectively. (Middle) Simulated actual noisy PM - Actual noisy e-ages (blue dots) values around the trend line (red) with specific $tr = 0.5$ and $\sigma_p = .8$. The (green) 45° line represents the c-age of each individual. (Right) The values inferred by the EPM-CEM algorithm - green dots represent the inferred e-age by the algorithm. It should be compared to the real e-age (blue). While there is a gap, linear with time, between actual and inferred e-ages, the trend is captured.

In order to simulate realistic e-ages, we allow for each individual, some stochastic deviation of her/his e-age from the (or population's) trend line, and that deviation depends on some variance σ_p. To show that, we set a specific tr and σ_p. Figure 2(M) shows simulated e-ages around the trend function with specific trend $tr = 0.5$ and $\sigma_p = .8$. For illustration, the straight 45° line, representing the c-age, appears in green the figure.

Finally, for every such e-age produced, our simulation procedure generates the methylation status $\hat{s}_{ij}$ for every site i and and individual j, according to the model: $\hat{s}_{ij} = s_i^0 + r_i t_j + \varepsilon_{i,j}$. Figure 2(R) shows the result of applying the EPM-CEM procedure to such synthetic data. Green dots represent the inferred e-age by the algorithm and should be compared to the real (model) e-age (blue, same as in the middle box). We can see that EPM-CEM is capable of capturing the trend imposed by the simulation however it lags below the trend line by a gap that is linearly (inversely) correlated with age (red-green dots). This gap, returned by the procedure is due to the degeneracy of the likelihood surface that allows for multiple points in the surface to attain the same likelihood and in particular the maximum likelihood.

Analysis of Human Data

We have shown above the ability of our technique to identify the correct trend in aging in simulated data, and we now move to analyze human methylation data from various types of data sets. The general procedure we have taken to identify and assess a trend is as follows. We applied the EPM approach to the real methylation data, all taken from the Gene Expression Omnibus (GEO) repository, using the procedure in the simulation section above **??**. The EPM, allows us to determine whether the MC hypothesis is rejected by the pacemaker, and also infers for each individual its epigenetic age. We remark that for all the real datasets we analyzed here, the pacemaker hypothesis was found superior to the linear approach with p-values always smaller than 10^{-6}. As these information is not essential to the main subject of this study, it appears in details in the supplementary text.

Similarly to the simulation study, in a subsequent stage to the EPM, we plot for each individual its two ages. Here however, as these points were not synthetically generated by a function, we attempt to fit a trend function best describing these points. We focus on two families of functions - linear and logarithmic, as they are very general with a single explaining variable. These families are indeed the most common for trend approximation. However, to obtain additional insights on the functional form of the trend, we also accompany the logarithmic and the linear approximations with a quadratic best approximation line. In the results below, we demonstrate how we exploit the added information provided by the quadratic approximation. We note that any linear line is a special case of a quadratic family, simply with quadratic coefficient equals zero, and hence by definition its fit is always inferior to the quadratic approximation.

Our analysis is divided to age range based analysis, and also to tissue based analysis. All data sets required a preprocessing step of selecting the most informative (1000) sites, and based on our conclusions above, we used the Pearson correlation coefficient (PCC) criterion.

5.3.0.1 Epigenetic Aging in Adults

Our first data set is from GSE40279, consisting of 656 blood samples from adults [11]. Their ages range from 19 years to 101. The results are depicted in

In the top two graphs of Figure 3 we show scatter plots with the three trend lines where the one with the best fit appears in the left and the two suboptimal ones in the right scatter plot. For each trend line, we also depict its exact formula, the R^2 and the adjusted R^2 for polynomial trend lines. It is quite evident that the linear trend is superior here to the other two, with negligible increase in R^2 for the quadratic trend. Therefore, in order to check if there is still a trend of non linearity, we applied the Mann-Kendall test as described in the Methods section. The lower part of Figure 3 shows the values of ρ_i ordered by chronological age. We set to test if there is an increasing trend in this series. The value obtained for S was 1619, under this size of data, we have a huge variance with $Var(S) = 31438253.33$ yielding z-score of 0.28856 which is not significant.

Our second data set is the GSE87571, also from human blood taken from 366 individuals of ages from 14 to 94 years old. The results of applying the EPM-ECM to this data are depicted inFigure 4. The top two graphs show the scatter plot of e-age versus c-age with the three trend types - linear, quadratic, and logarithmic. The difference between the linear and the quadratic is negligible, with a $R^2 = 0.612$ and $R^2 = 0.613$ respectively. Nevertheless, we see a bend in the points corresponding to younger ages. In general, the entire collection of points here allude to a concave shape, i.e. a decreasing function, as can be seen by the negative coefficient (-0.0045) of the quadratic term in the quadratic trend line. This should be contrasted to the convex trend of the quadratic trend in the previous case (Figure 3) where the first coefficient equals 0.0024.

To verify this decreasing trend, as in the previous data set, we apply the test of monotonicity - the Mann-Kendall Test - to this data. The value obtained for S was 4225 implying that we have an increasing trend in ρ and therefore epigenetic aging is decreasing in time also

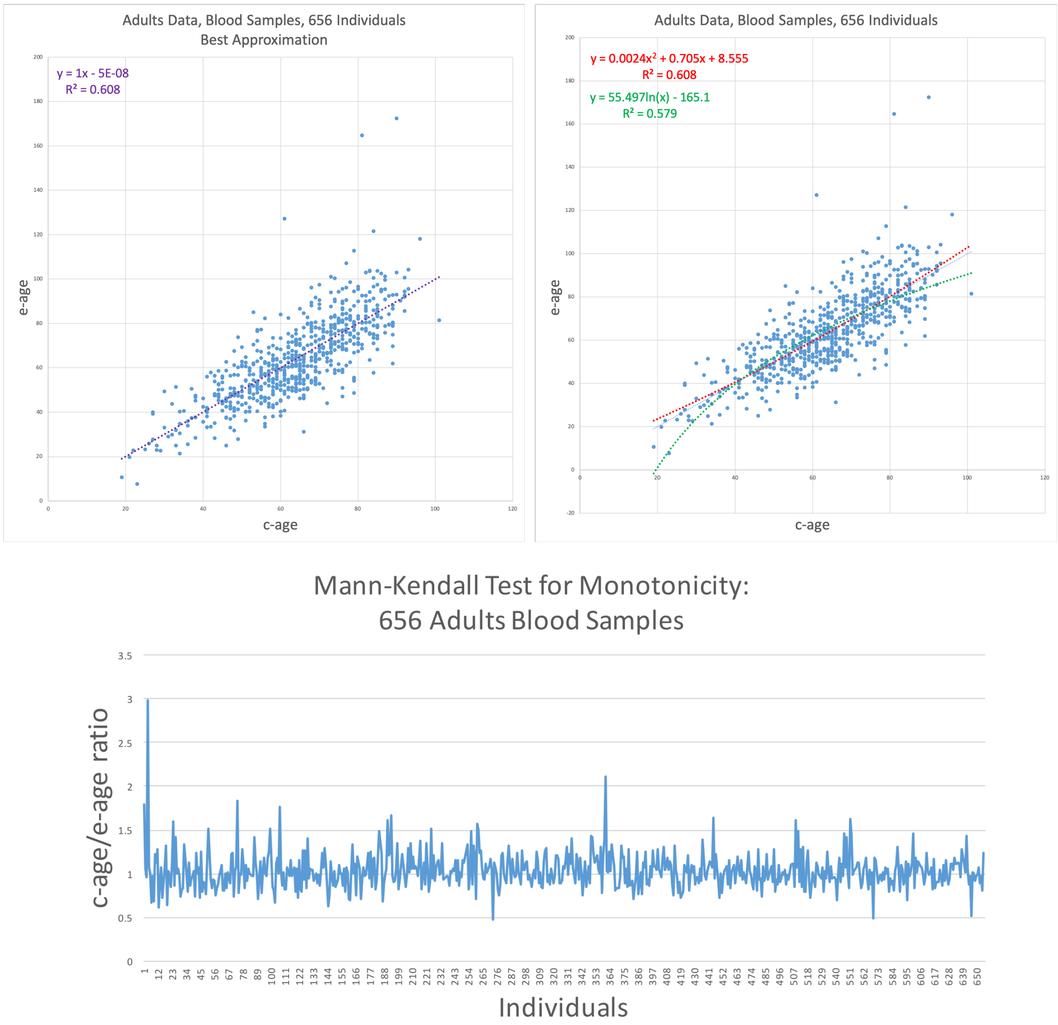


Figure 3: GSE40279 - Human Blood Data Results I

(Top) e-age vs c-age in adults. Age is plotted in years. The left graph shows the best approximation to the data. The linear line is slightly and insignificantly inferior to the quadratic approximation and therefore is the best fit.(Bottom) Mann-Kendall Test for monotonicity Trend: c-age vs e-age ratio ordered from left to right according to c-age. If rate of aging is decreasing, we expect to see monotonic increase in the function. Indeed the function is increasing but not in a significant manner.

for this data set. The variance here is $Var(S) = 5469768$ yielding a z-score of 1.806 and a p-value smaller than 0.04 and is therefore significant.

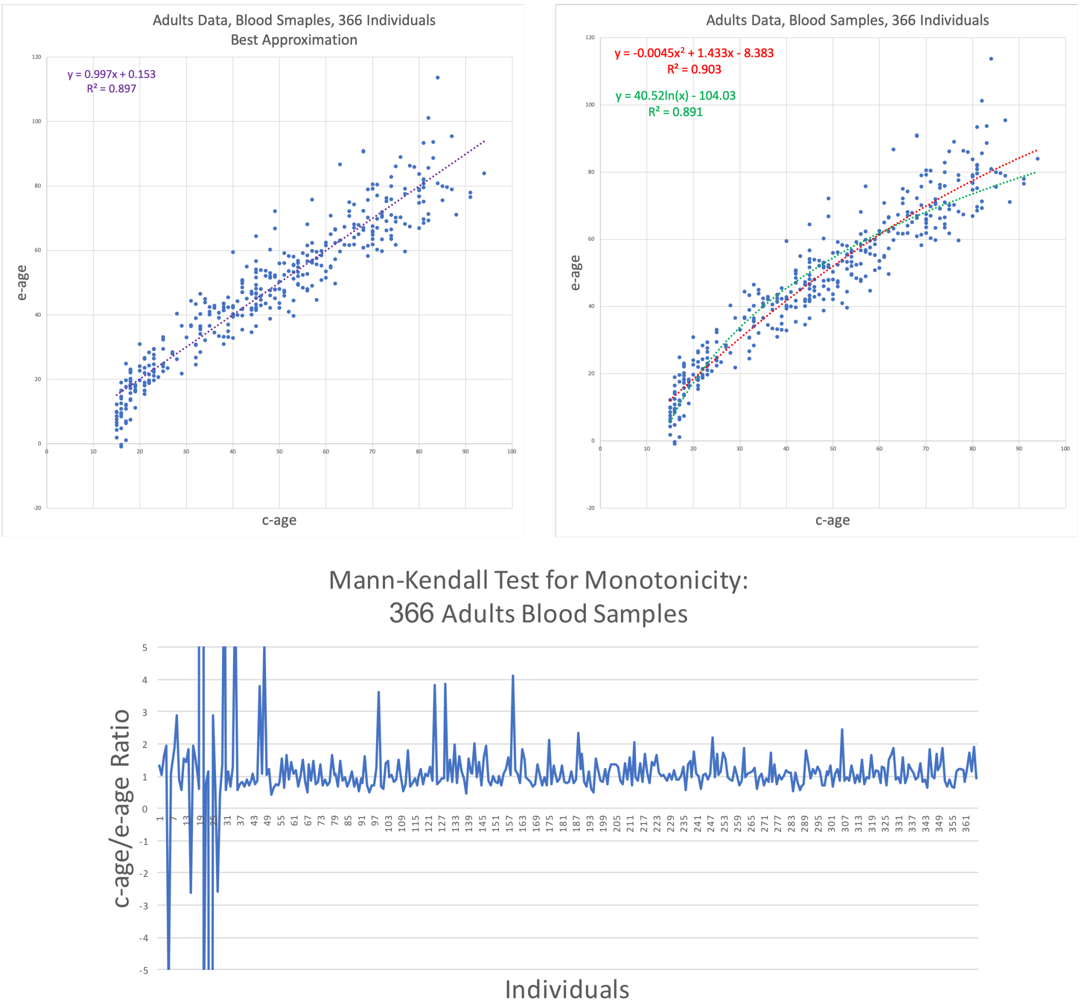


Figure 4: GSE87571 - Human Blood Data Results II

(Top) e-age vs c-age in adults. Age is plotted in years. The left graph shows the best approximation to the data. The linear line is slightly and insignificantly inferior to the quadratic approximation. (Bottom) Mann-Kendall Test for monotonicity Trend: c-age vs e-age ratio ordered from left to right according to c-age. If rate of epigenetic aging is decreasing with time, we expect to see a monotonic increase in the $\frac{\text{c-age}}{\text{e-age}}$-ratio function. Indeed the function is significantly increasing.

Epigenetic Aging in Children

After analyzing the data collected from adults, we turned to analyze data from children. We analyze the GSE36064 data set of blood samples taken from 78 children of ages ranging from one year to 16. The ages here, as well as in the figure describing it, are represented in months. The results are shown in Figure 5. Here, the linear trend is clearly inferior to the quadratic and the logarithmic trends. We find that the logarithmic trend is the best approximation and show it in the left side of Figure 5.

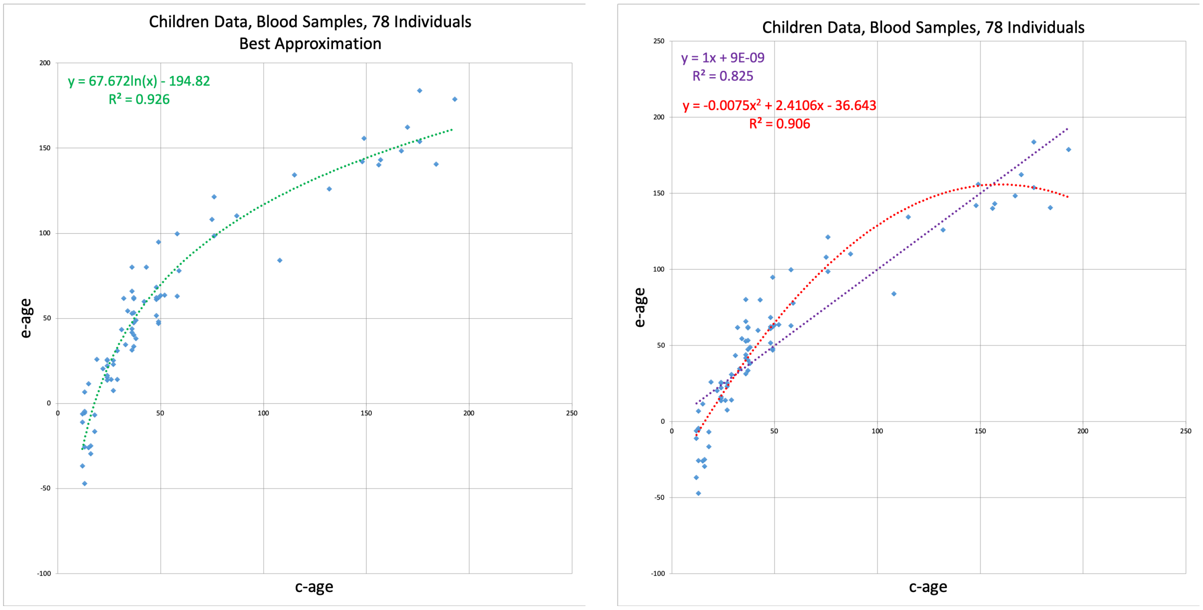


Figure 5: GSE36064 - Children Blood Data Results

e-age vs c-age in young humans. Age is plotted in months. The left graph shows the best approximation to the data. The logarithmic approximation provides the best explanation.

Combined Age Analysis

In the previous data sets we restricted the analysis to specific age ranges, such as children or adults. In the next two data sets we analyze blood samples from individuals with age ranges from childhood to old age. The first data set - GSE60132 - was taken from peripheral blood samples of 192 individuals of Northern European ancestry [2]. Ages range from 6 to 85 years. The results are shown in Figure 6. As can be noticed, the logarithmic trend line provides better R^2 than the linear trend line, 0.912 versus 0.899. The concavity of the spread

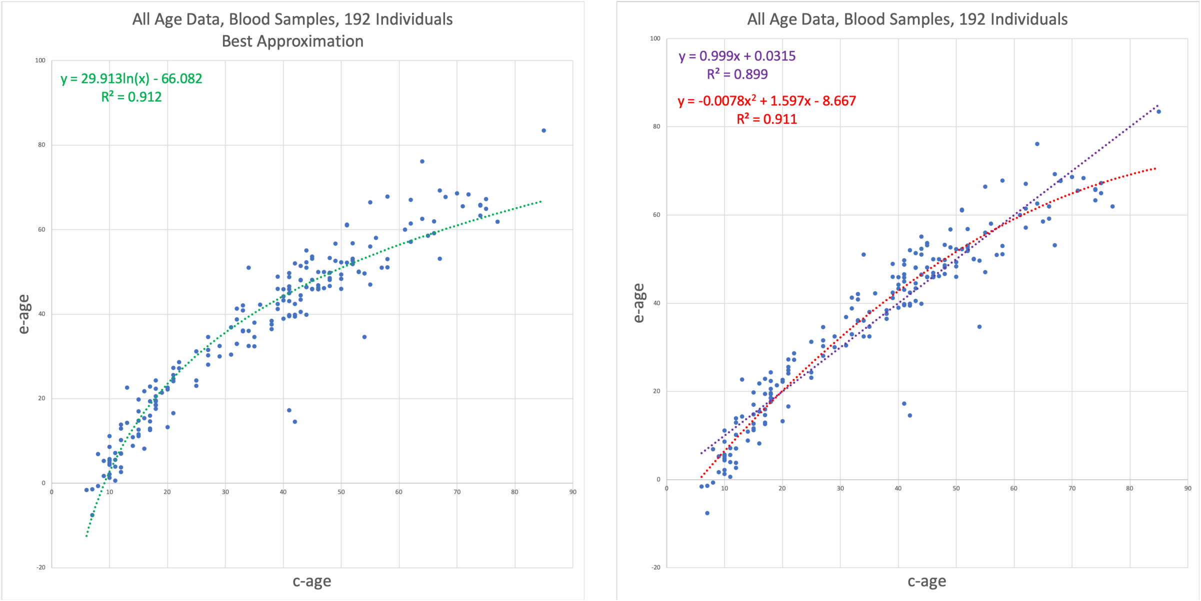


Figure 6: GSE60132 - Human, All Ages, Blood Data Results I

e-age vs c-age in a wide age range. Age is plotted in years. The left graph shows the best trend line approximation to the data, which is the logarithmic trend function. At the right, the inferior trends - the quadratic and linear. The quadratic line is slightly and insignificantly inferior to the logarithmic approximation, buy also portrays a concave line due to negative first coefficient -0.0078.

of the points is fairly noticeable and this is confirmed by the negative first coefficient of the quadratic trend function - -0.0078.

The Next data set - GSE64495 - is also from blood samples of 113 individuals [31]. Here, while there is a scarcity of samples from the age range 12-35, the entire age range of the study begins at even younger ages than the previous data set: 2.3 years versus 6. Our results for this dataset are depicted in Figure 7. Here the advantage of the logarithmic trend line over the linear is the most significant among the adults containing data sets analyzed so far, $R^2 = 0.924$ versus $R^2 = 0.866$, and is even significant over the quadratic - $R^2 = .902$. The decrease in the rate is evident as well as the fit to the logarithmic trend line.

Brain Development and Aging

Our last data set is from GSE74193, consisting of 675 samples from brain tissues from before birth to old age [**Jaffe-Natneuro-2016**]. The advantage of this data set is two-fold. First, the broad range of ages - from half a year before birth to 85 years, which represents a broader age range than that found in the the previous data sets, and allows us to track epigenetic aging across the entire span of life, starting from before birth. Second, all the samples from the previously analyzed data sets came from blood. This data set, from brain tissues, allows us to contrast our results from blood tissues to another tissue type. Our results are depicted in Figure 8. The logarithmic approximation, appears on the left graph, not only provides a significantly better fit to the data, with $R^2 = 0.975$ versus $R^2 = 0.864$ and $R^2 = 0.707$ for the quadratic trend line and the linear trend line respectively. Moreover, the high R^2

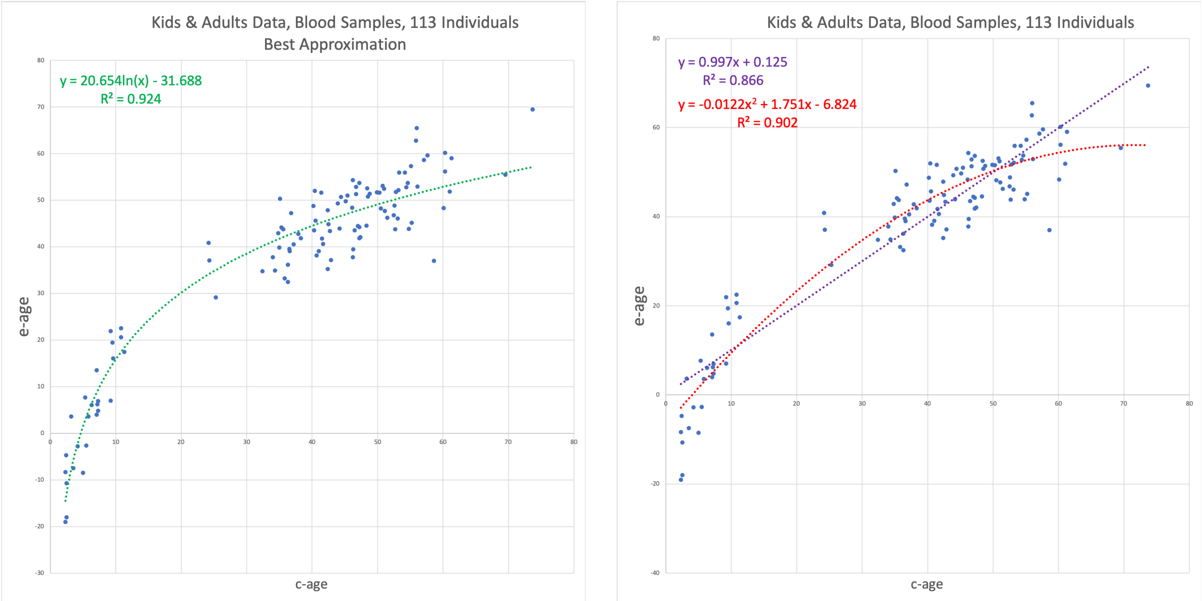
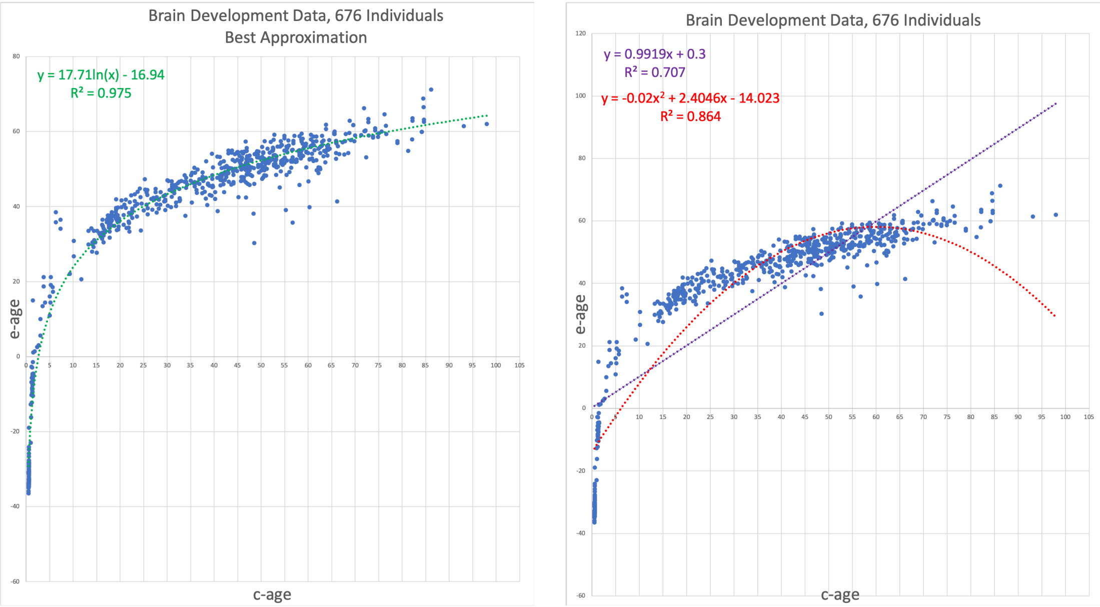


Figure 7: GSE64495 - Human, All Ages, Blood Data Results II

. e-age vs c-age in kids and adults. Age is plotted in years. The left graph shows the best approximation to the data, obtained by the logarithmic trend line with $R^2 = 0.924$. On the right, the inferior trend lines, the linear line with $R^2 = 0.866$ and the quadratic line with $R^2 = 0.902$.

provides an almost perfect fit to the data.



Brain Development GSE74193 - Brain Development// e-age vs c-age in young humans. Age is plotted in years. The left graph shows the best approximation to the data. The logarithmic approximation provides the best explanation.

EPM Comparison to the Horvath and Hannum Epigenetic Clocks

Next we wanted to evaluate the EPM in context of the well known Horvath [13] and Hannum [11] epgienetic clocks. To generate sufficiently large experimental data we combined Human Illumina methylation 450K Beadchip data generated using whole blood across several experiments [3, 22, 14, 30, 15, 5, 9, 17, 6] from GEO. To facilitate cross experiment comparisons, we performed stratified quantile normalization for different probe technologies used in the Illumina 450k Beachip Array as previously reported [13]. Comparisons of the EPM model to the Horvath and Hannum models of epigenetic aging were performed as follows. The combined methylation data was subset to include CpG sites reported in the Horvath or Hannum models. Samples with missing data for any of the CpG sites were dropped. The methylation subsets consisted of 354 CpG sites and 614 samples and 71 CpG sites and 1117 samples for the Horvath and Hannum comparison datasets respectively. We then calculated the Horvath and Hannum epigenetic ages for each sample. The EPM produced a similar estimate of epigenetic age for both the Horvath ($R^2 = 0.901$) and Hannum ($R^2 = 0.957$) models when the min-max scaled [**scikit**] estimates of epigenetic age were compared to the scaled EPM ages. Both the Horvath and Hannum models showed a non-linear epigenetic aging trend as shown in Figure 9. However, the computed Horvath epigenetic age represents a log transformed age under 20 years and linear age thereafter. While the raw output of the Horvath clock displayed a better fit to a logarithmic epigenetic aging trend than the transformed Horvath epigenetic age estimate, $R^2 = 0.859$ vs $R^2 = 0.785$.

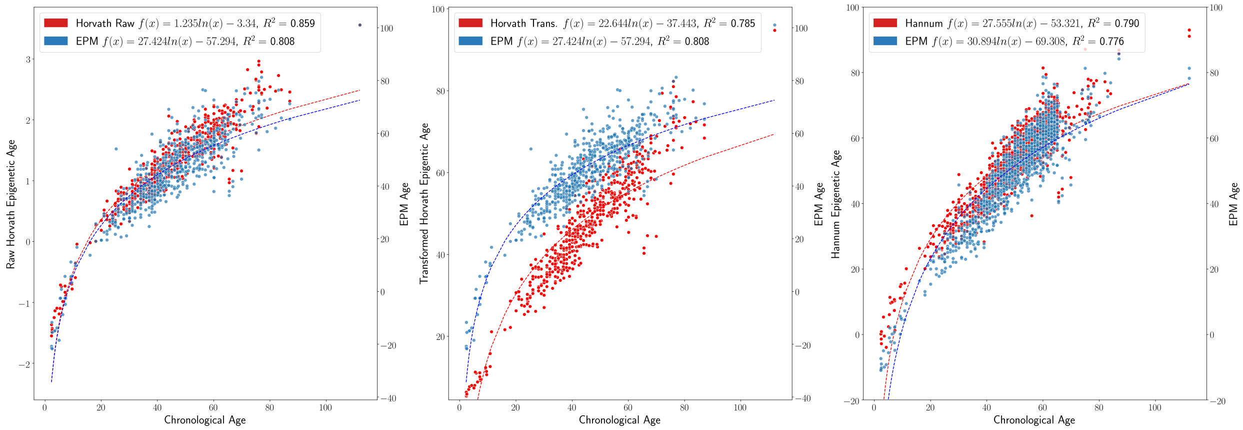


Figure 9: EPM Hannum Horvath Trend Comparison
(Left) EPM and Horvath aging trend. (Middle) EPM and Horvath (transformed ages) aging trend. (Right) EPM and Hannum aging trend.

To assess the relative importance of individual CpG sites in the Horvath and Hannum models to the EPM, we compared the EPM rate for each CpG site to the respective regression coefficient in Horvath and Hannum models as shown in Figure 10. EPM rates were correlated with the regression coefficients in both the Horvath ($\rho = 0.565$, $p = 3.19e-31$) and Hannum ($\rho = 0.519$, $p = 3.4e-6$) models. Interestingly, the relative importance of many CpG sites appears to differ between the EPM and Horvath and Hannum models; particularly in the case of the Horvath model. Many of the CpG sites assigned high coefficients in the Horvath model have a rate close to zero in the EPM model and sites with large absolute EPM rates have coefficients near zero in the Horvath model. This trend appears to be associated with the methylation state of the individual CpG sites. Hypomethylated sites are assigned coefficients with greater magnitude in the Horvath model, while intermediate sites ($0.3 < \beta < 0.7$) are assigned higher magnitude rates in the EPM model. A similar relationship isn't observed when comparing the Hannum model coefficients and EPM rates.

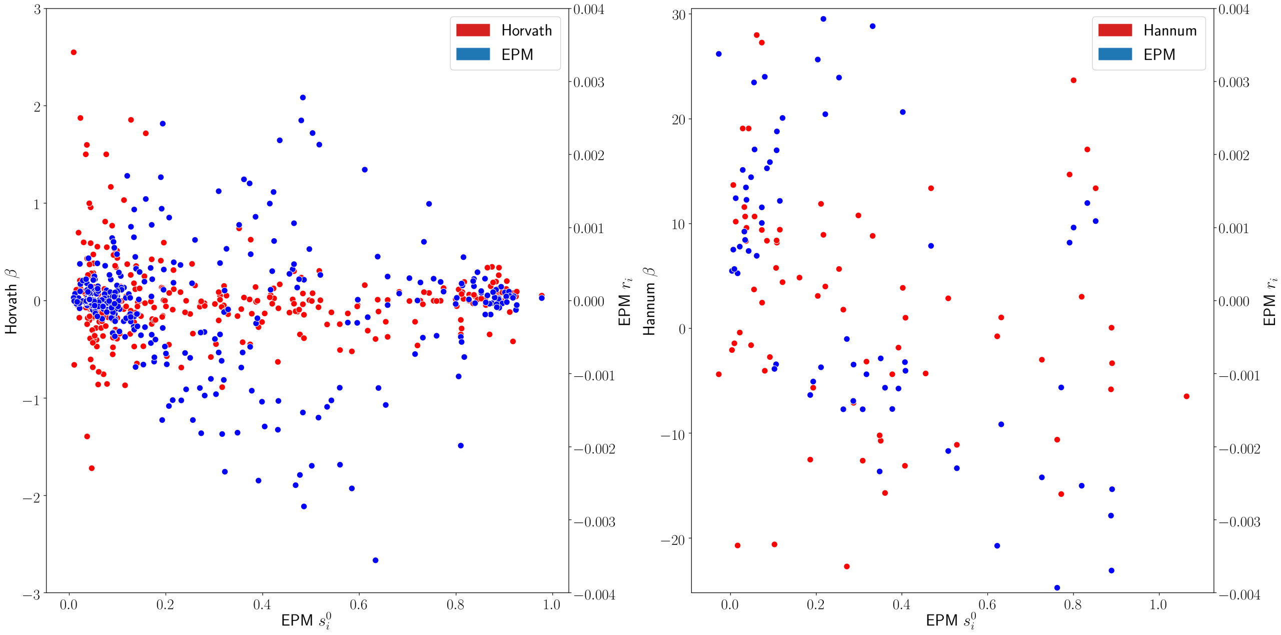


Figure 10 EPM Hannum Horvath Rate Coefficient Comparison

Horvath and Hannum site coefficients compared to the EPM initial methylation values s_i^0 and EPM rates r_i for the CpG sites used in the Horvath and Hannum models respectively. (Left) EPM and Horvath (Right) EPM and Hannum

5.4 Discussion

During the past few years several studies have shown that DNA methylation patterns continue to change as individuals age. These observations have been leveraged to construct epigenetic clocks that predict the age of an individual based on their methylation profile. While these tools have proven to be very useful for aging studies, they are based on a priori assumptions about the relationships between epigenetic and chronological age. For example, the Hannum age clock assumes that epigenetic age and chronological are linearly related, and using multivariate penalized regression identifies 71 CpG sites whose methylation values can be combined using a weighted sum to predict the actual age of the individual. The Horvath clock uses a more complex set of assumptions to derive its predictions of chronological age: it applies a logarithmic transformation to ages below 20 and a linear relationship for ages greater than 20. He then identifies 354 CpG sites using elastic net regression to very accurately predict the transformed age. These biomarkers have been widely used to study aging, as is evident by the hundreds of studies that have utilized them. However, because of their underlying assumptions, they are not ideal tools to infer trends in epigenetic aging rates across life spans in an unbiased fashion. Nevertheless, correctly modeling potential nonlinearities in epigenetic aging is critical to advance the field and generate an even more robust understanding of epigenetic aging and its impact on human health and mortality.

To address this question, we have developed an unbiased approach to measure the trends in epigenetic aging within a cohort of individuals of varied ages. Our approach is distinct from the Hannum and Horvath epigenetic clocks in that it is not designed to optimally predict the age of an individual, but rather to model the non-linear trends in epigenetic aging over time without making any a priori assumptions about what these may be. The method is inspired by evolutionary models that attempt to model mutation rate changes over time. This lead to the development of our epigenetic universal pacemaker model, which we have presented in previous studies [27, 26].

To attempt to identify rates of epigenetic aging over the entire life span of humans we apply the epigenetic pacemaker to multiple datasets that measure DNA methylation in large cohorts of individuals of varying ages. In the first few cohorts the methylation is profiled from blood, while in the last cohort the methylation is measured in brain tissue. In all cases methylation is profiled using an Illumina microarray that measures the methylation across approximately 450,000 sites. While all cohorts sample individuals from early adulthood to old age, the second set also include samples from individuals from early childhood to adulthood, and in the last case of the brain study, also samples from fetuses obtained before birth.

By applying our epigenetic pacemaker model to this data we observe consistent and robust trends across these datasets. The first is that from early adulthood (around age 20) to old age (well into the 90s), DNA methylation changes in a roughly linear fashion, yet with a slight but significant tendency for rate decrease. However, in contrast to the adults, we find that DNA methylation changes are strongly nonlinear from late fetal stages to adolescents. In both the blood and brain datasets that measure these stages we observe that the epigenetic age inferred by our model is related to the chronological age by a logarithmic transformation across the entire span of life, from before birth to old age. Thus, DNA methylation changes are very rapid initially, and then gradually decrease with age. This implies that the rate of change of epigenetic ages (i.e. the slope of our trend line) is roughly the inverse of the chronological age. The fact that we consistently observed these trends across multiple datasets, and two different tissues, suggests that the logarithmic relationship between epigenetic and chronological age may be a universal property of human aging.

This universal logarithmic trend may help explain some interesting observations that have been reported in the literature regarding epigenetic aging. For example, one recent study found that the Horvath epigenetic clock ”systematically underestimates ages in tissues from older people,” and that ”a decrease in slope of the predicted ages were observed at approximately 60 years, indicating that some loci in the model may change differently with age, and that age acceleration measures will themselves be age-dependent"[7]. A second study also found that "epigenetic age increases at a slower rate than chronological age across the life course, especially in the oldest population"[21]. These results suggest that the underlying assumptions about the relationships between epigenetic age and chronological age impact the performance of the Hannum and Horvath epigenetic clocks, and that deviations are most notable in old age. We speculate that if the logarithmic epigenetic aging trend that we observe is in fact universal, that this could lead to improved biomarkers that show more robust performance at the extremes of the age distributions, leading to more accurate associations between epigenetic aging and human health and longevity.

Moreover, we believe that the observation that epigenetic aging is logarithmic over the entire life span opens up new avenues for epigenetic research in the future. What mechanisms lead to the gradual reduction in epigenetic rates from late fetal stages to centenarians? Is the logarithmic trend related to prior observations that epigenetic aging is a measure of epigenetic entropy? The answers to these questions will undoubtedly influence our understanding of human aging and longevity and will most likely apply to a broad range of organisms. By quantitatively demonstrating these trends in an unbiased fashion, we believe we have laid a solid foundation for the development of improved aging biomarkers and the investigation of the underlying mechanisms of epigenetic aging, and ultimately to the answers to these important questions that are fundamental to the biology of development and aging.

In response to our first question about mechanisms of epigenetic age we speculate that epigenetic aging is partially driven by entropic forces. Under this assumption, epigenomes are set early in life and during aging the methylation levels of promoters and enhancers drift away from their original state towards a level of intermediate methylation, which represents more disordered states. Thus, we hypothesize that if methylation profiles were generated at the single cell level, we would find that ensembles of cells of the same type are more epigenetically similar early in life than later in life. Moreover, we speculate that the logarithmic trend between epigenetic age and time is reminiscent of the relationship between entropy and the number of states accessible to a system, that are also related by a logarithmic relationship in Boltzmann's formula. Thus, if we hypothesize that Epigenetic age is a measure of entropy, then we would conclude that the number of epigenetic states accessible to an individual increase linearly in time.

Supplementary Text

Solving the MC Model

Under the statistical framework defined above minimizing RSS is equivalent to maximizing the likelihood function L. In particular the ML RSS, $\widehat{RSS}$, is used for computing χ^2. RSS is a polynomial over the variables r_i and s_i^0 where every monomial in the RSS stands for an entry in the input matrix $\hat{S}$, that is $\hat{s}_{i,j}$, and is of the form:

$$\varepsilon_{i,j}^2 = (\hat{s}_{i,j} - t_j r_i - s_i^0)^2, \tag{5.3}$$

where in our case the inputs are the $\hat{s}_{i,j}$ and t_j and the variables sought are r_i and s_i^0, for every $i \leq n$ (our set of sites).

Normally, critical points of the RSS are found through partial derivatives of the RSS with respect to every such variable. The critical points are the points in the $2n$ space where all these partial derivatives simultaneously vanish [28]. Finding these points is normally carried out using some numerical method.

In our case however, the special structure of the problem allows us a more efficient solution.

When the residuals are linear in all unknowns, a solution can be found using linear algebra tools which have a closed form solution (given that the columns of the matrix are linearly independent). Under this formalization the optimal (ML) solution is given by the vector $\hat{\beta}$ as follows:

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T y, \tag{5.4}$$

where X is a matrix over the variable's coefficients in the problem, y is a vector holding the observed values - in our case the entries of $\hat{S}$, and the RSS equation can be written such that for every row i in X, $y_i - \sum_j X_{i,j}\beta_j$ is a component in the RSS. Equation (5.4) is solved precisely by standard linear algebra operations, however quite computation intensive for the sizes of our problem. Recall that for m individuals and n sites, our RSS contains mn components each of which corresponds to an entry in $\hat{S}$ in the form $\hat{s}_{i,j} - t_j r_i - s_i^0$ where $\hat{s}_{i,j}$ and t_j are input parameters. This leads to the following observation (stated in [27]): [27] Let X be a $mn \times 2n$ matrix whose kth row corresponds to the (i,j) entry in S, the first n variables of β are the r_i's and the second n variables are the s_i^0's, and the $im+j$ entry in y contains $s_{i,j}$ (see Figure S1). Then, if we set the k row in X all to zero except for t_j in the i'th entry of the first half and 1 in i'th entry of the second half, we obtain the desired system of linear equations (see again illustration for row setting in Figure S1).

The likelihood score is calculated by plugging in the values obtained for $\hat{\beta}$ in (5.4) to the likelihood function (or alternatively into the RSS).

Likelihood Ratio Test

The likelihood ratio test (LRT) [33] is a statistical test used to compare the goodness of fit of two competing models, one of which (the null model) is a special case of the other, more general, one. The log of the ratio of the two likelihood scores distributes as a χ^2 statistic (where the degree of freedom (DF) is the difference in the number of free variables between the two models) and therefore can be used to calculate a p-value. This p-value is used to reject the null model in the conventional manner. Specifically, let $\Lambda = L_0/L_1$ where L_0 and L_1 are the maximum likelihood values under the restricted and the more general models respectively. Then asymptotically, $-2\log(\Lambda)$ will distribute as χ^2 with degrees of freedom equal the number of parameters that are lost (or fixed) under the restricted model.

In our case, we first set the two RSS, $\widehat{RSS}_{MC}$ and $\widehat{RSS}_{PM}$ for the ML values for RSS under

$$\mathcal{X} = \left[\begin{array}{ccccccccc} t_1 & 0 & \cdots & 0 & | & 1 & 0 & \cdots & 0 \\ t_2 & 0 & \cdots & 0 & | & 1 & 0 & \cdots & 0 \\ & & \vdots & & | & & \vdots & & \\ t_m & 0 & \cdots & 0 & | & 1 & 0 & \cdots & 0 \\ 0 & t_1 & \cdots & 0 & | & 0 & 1 & \cdots & 0 \\ 0 & t_2 & \cdots & 0 & | & 0 & 1 & \cdots & 0 \\ & & \vdots & & | & & \vdots & & \\ 0 & t_m & \cdots & 0 & | & 0 & 1 & \cdots & 0 \\ & & \vdots & & | & & \vdots & & \\ & & \vdots & & | & & \vdots & & \\ 0 & \cdots & 0 & t_1 & | & 0 & \cdots & 0 & 1 \\ 0 & \cdots & 0 & t_2 & | & 0 & \cdots & 0 & 1 \\ & & \vdots & & | & & \vdots & & \\ 0 & \cdots & 0 & t_m & | & 0 & \cdots & 0 & 1 \end{array}\right] \begin{bmatrix} r_1 \\ \vdots \\ r_n \\ - \\ s_1^0 \\ \vdots \\ s_n^0 \end{bmatrix} = \begin{bmatrix} \hat{s}_{1,1} \\ \hat{s}_{1,2} \\ \\ \vdots \\ \\ \hat{s}_{n,m} \end{bmatrix} \tag{5.5}$$

Figure S1: The $mn \times 2n$ matrix X that is used in our closed form solution to the MC case. Every row corresponds to a component in the RSS polynomial and the corresponding entries (ith and $i+n$th) in that row are set to t_j and 1 respectively.

MC and PM respectively as obtained by the algorithm.

$$RSS = \sum_{i \le n} \sum_{j \le m} (\hat{s}_{i,j} - (s_i^0 + r_i t_j))^2. \tag{5.6}$$

Now, it is easy to see that

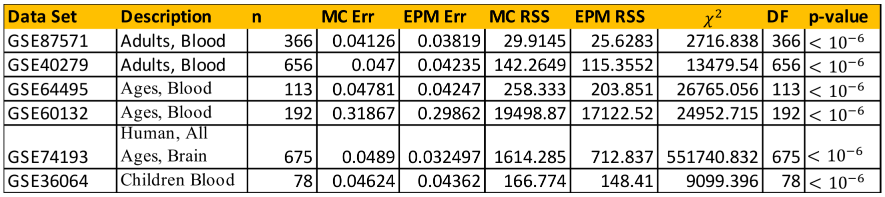
$$\log(\Lambda)) = -\frac{nm}{2} \log \frac{\widehat{RSS}_{MC}}{\widehat{RSS}_{PM}}. \tag{5.7}$$

Hence we set our χ^2 statistic as

$$\chi^2 = nm \log\left(\frac{\widehat{RSS}_{MC}}{\widehat{RSS}_{PM}}\right). \tag{5.8}$$

Detailed Statistics of the Datasets' Analysis

The following table provides statistics for each of the datasets analysed in the Results section in the main text. As can be seen, all data sets attained p-value significantly smaller than 10^{-6} implying the pacemaker hypothesis is significantly superior to the molecular clock.



Data Set	Description	n	MC Err	EPM Err	MC RSS	EPM RSS	χ^2	DF	p-value
GSE87571	Adults, Blood	366	0.04126	0.03819	29.9145	25.6283	2716.838	366	$< 10^{-6}$
GSE40279	Adults, Blood	656	0.047	0.04235	142.2649	115.3552	13479.54	656	$< 10^{-6}$
GSE64495	Ages, Blood	113	0.04781	0.04247	258.333	203.851	26765.056	113	$< 10^{-6}$
GSE60132	Ages, Blood	192	0.31867	0.29862	19498.87	17122.52	24952.715	192	$< 10^{-6}$
GSE74193	Human, All Ages, Brain	675	0.0489	0.032497	1614.285	712.837	551740.832	675	$< 10^{-6}$
GSE36064	Children Blood	78	0.04624	0.04362	166.774	148.41	9099.396	78	$< 10^{-6}$

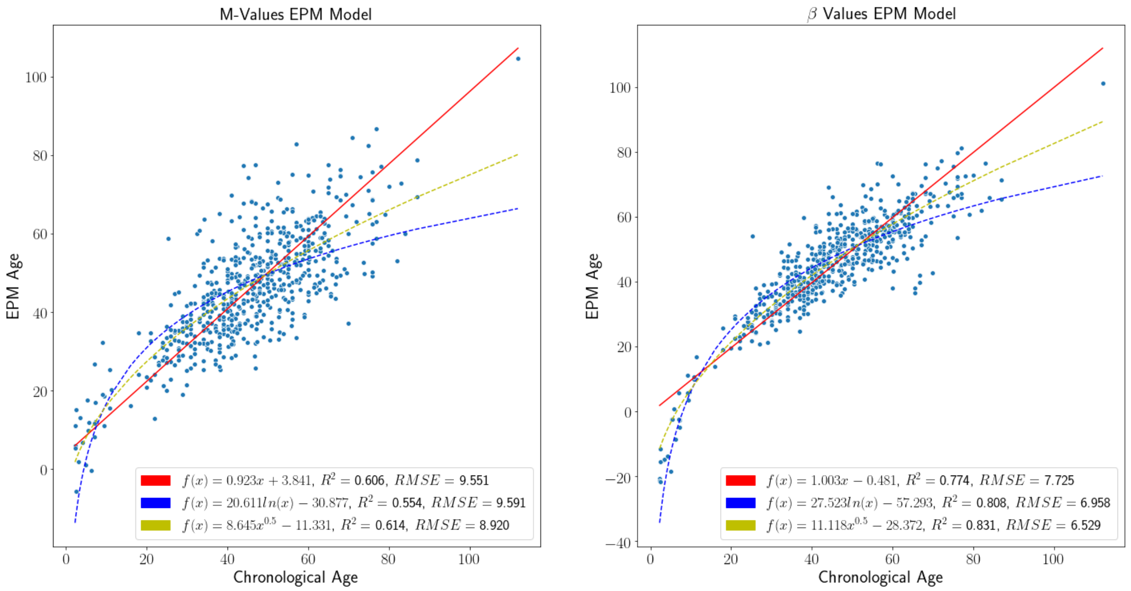
S.Figure 2: Datasets error statistics and p-values for rejecting the molecular clock hypothesis.

EPM Model Fit, M-Values to β-Value Comparison

We evaluated the performance of the EPM model on a matrix of M-values by logit transforming, $M_i = log_2(\frac{\beta_i}{1-\beta_i})$, all beta values in the Horvath comparison methylation dataset. A five-fold cross validated EPM model was then fit using the matrix of M-values and using a matrix of beta values. The fit trend lines for the M-value EPM performed worse at modeling aging over time than the fit lines to the beta value EPM by R^2 and root mean squared error (RMSE), see Figure S2.

Functional Annotation of Horvath, EPM Comparison

We explored the relationship between the Horvath coefficients and EPM rates of the same CpG sites by annotated each site with function information previously reported [18]. We were unable to find a clear connection between CpG Horvath site coefficients and EPM rates, see S.Figure 4.



S.Figure3: EPM Model, M-values (Left) and β-Values (Right).

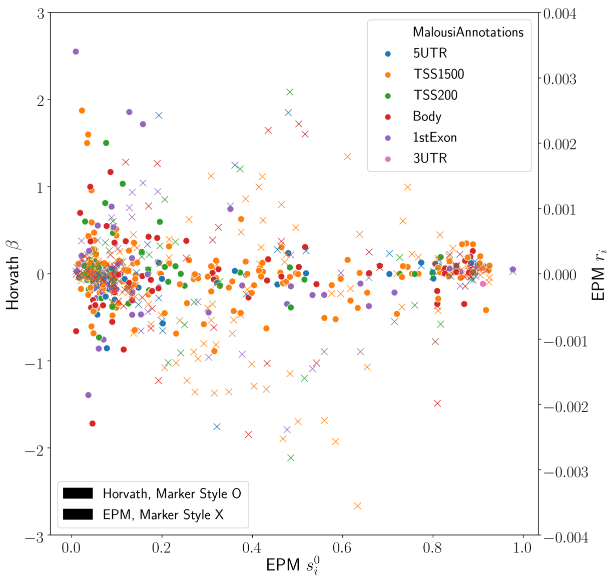


Figure S.4: EPM rates, Horvath coefficients, annotated with functional information from Malousi et al. 2018.

Bibliography

[1] Chen et al. "DNA methylation-based measures of biological age: meta-analysis predicting time to death". In: *Aging (Albany NY)* 8.9 (Sept. 2016), pp. 1844–1859. DOI: 10.18632/aging.101020. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5076441/.

[2] Omar Ali et al. "An epigenetic map of age-associated autosomal loci in northern European families at high risk for the metabolic syndrome". In: *Clinical Epigenetics* 7.1 (Feb. 2015), p. 12. ISSN: 1868-7083. DOI: 10.1186/s13148-015-0048-6. URL: https://doi.org/10.1186/s13148-015-0048-6.

[3] J T Bell et al. "Differential methylation of the TRPA1 promoter in pain sensitivity." eng. In: *Nature communications* 5 (2014), p. 2978. ISSN: 2041-1723 (Electronic). DOI: 10.1038/ncomms3978.

[4] Bradley E. Bernstein, Alexander Meissner, and Eric S. Lander. "The Mammalian Epigenome". In: *Cell* 128.4 (2007), pp. 669–681. ISSN: 0092-8674. DOI: http://dx.doi.org/10.1016/j.cell.2007.01.033. URL: http://www.sciencedirect.com/science/article/pii/S0092867407001286.

[5] Katrine Braekke Norheim et al. "Epigenome-wide DNA methylation patterns associated with fatigue in primary Sjogren's syndrome." eng. In: *Rheumatology (Oxford, England)* 55.6 (June 2016), pp. 1074–1082. ISSN: 1462-0332 (Electronic). DOI: 10.1093/rheumatology/kew008.

[6] Aldo Cordova-Palomera et al. "Epigenetic outlier profiles in depression: A genome-wide DNA methylation analysis of monozygotic twins." eng. In: *PloS one* 13.11 (2018), e0207754. ISSN: 1932-6203 (Electronic). DOI: 10.1371/journal.pone.0207754.

[7] Louis Y El Khoury et al. "Properties of the epigenetic clock and age acceleration". In: *bioRxiv* (2018). DOI: 10.1101/363143. eprint: https://www.biorxiv.org/content/

early/2018/07/06/363143.full.pdf. URL: https://www.biorxiv.org/content/early/2018/07/06/363143.

[8] Suhua Feng, Steven E Jacobsen, and Wolf Reik. "Epigenetic reprogramming in plant and animal development". In: *Science* 330.6004 (2010), pp. 622–627.

[9] Eduardo Fernandez-Rebollo et al. "Primary Osteoporosis Is Not Reflected by Disease-Specific DNA Methylation or Accelerated Epigenetic Age in Blood." eng. In: *Journal of bone and mineral research : the official journal of the American Society for Bone and Mineral Research* 33.2 (Feb. 2018), pp. 356–361. ISSN: 1523-4681 (Electronic). DOI: 10.1002/jbmr.3298.

[10] R.O. Gilbert. *Statistical Methods for Environmental Pollution Monitoring*. Wiley, 1987.

[11] Gregory Hannum et al. "Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates". In: *Molecular Cell* 49.2 (2013), pp. 359–367. ISSN: 1097-2765. DOI: http://dx.doi.org/10.1016/j.molcel.2012.10.016. URL: http://www.sciencedirect.com/science/article/pii/S1097276512008933.

[12] Gary C Hon et al. "Epigenetic memory at embryonic enhancers identified in DNA methylation maps from adult mouse tissues". In: *Nature Genetics* 45 (Sept. 2013), 1198 EP -. URL: http://dx.doi.org/10.1038/ng.2746.

[13] Steve Horvath. "DNA methylation age of human tissues and cell types". In: *Genome Biology* 14.10 (2013), pp. 1–20. ISSN: 1474-760X. DOI: 10.1186/gb-2013-14-10-r115. URL: http://dx.doi.org/10.1186/gb-2013-14-10-r115.

[14] Steve Horvath and Andrew J. Levine. "HIV-1 infection accelerates age according to the epigenetic clock". In: *Journal of Infectious Diseases* (2015). DOI: 10.1093/infdis/jiv277. eprint: http://jid.oxfordjournals.org/content/early/2015/05/11/infdis.jiv277.full.pdf+html. URL: http://jid.oxfordjournals.org/content/early/2015/05/11/infdis.jiv277.abstract.

[15] Steve Horvath et al. “The cerebellum ages slowly according to the epigenetic clock.” eng. In: *Aging* 7.5 (May 2015), pp. 294–306. ISSN: 1945-4589 (Electronic).

[16] Peter A. Jones and Daiya Takai. “The Role of DNA Methylation in Mammalian Epigenetics”. In: *Science* 293.5532 (2001), pp. 1068–1070. ISSN: 0036-8075. DOI: 10.1126/science.1063852. eprint: http://science.sciencemag.org/content/293/5532/1068.full.pdf. URL: http://science.sciencemag.org/content/293/5532/1068.

[17] Lara Kular et al. “DNA methylation as a mediator of HLA-DRB1*15:01 and a protective variant in multiple sclerosis.” eng. In: *Nature communications* 9.1 (June 2018), p. 2397. ISSN: 2041-1723 (Electronic). DOI: 10.1038/s41467-018-04732-5.

[18] Andigoni Malousi et al. “Age-dependent methylation in epigenetic clock CpGs is associated with G-quadruplex, co-transcriptionally formed RNA structures and tentative splice sites”. In: *Epigenetics* 13.8 (2018), pp. 808–821. ISSN: 15592308. DOI: 10.1080/15592294.2018.1514232. URL: https://doi.org/10.1080/15592294.2018.1514232.

[19] H.B. Mann. “Non-parametric tests against trend”. In: *Econometrica* 13 (1945), pp. 163–171.

[20] Riccardo E Marioni et al. “The epigenetic clock and telomere length are independently associated with chronological age and mortality”. In: *International Journal of Epidemiology* 47.1 (Feb. 2018), pp. 356–356. URL: http://dx.doi.org/10.1093/ije/dyx233.

[21] Riccardo E Marioni et al. “Tracking the Epigenetic Clock Across the Human Life Course: A Meta-analysis of Longitudinal Cohort Data”. In: *The Journals of Gerontology: Series A* (2018), gly060. DOI: 10.1093/gerona/gly060. eprint: /oup/backfile/content_public/journal/biomedgerontology/pap/10.1093_gerona_gly060/2/gly060.pdf. URL: http://dx.doi.org/10.1093/gerona/gly060.

[22] Dragan Milenkovic et al. “Dietary flavanols modulate the transcription of genes associated with cardiovascular pathology without changes in their DNA methylation state.”

eng. In: *PloS one* 9.4 (2014), e95527. ISSN: 1932-6203 (Electronic). DOI: 10.1371/journal.pone.0095527.

[23] Marco Morselli et al. “In vivo targeting of de novo DNA methylation by histone modifications in yeast and mouse”. In: *eLife* 4 (Apr. 2015). Ed. by Bing Ren, e06205. ISSN: 2050-084X.

[24] S. Snir, Y. I. Wolf, and E. V. Koonin. “Universal pacemaker of genome evolution in animals and fungi and variation of evolutionary rates in diverse organisms”. In: *Genome Biology and Evolution* (2014). DOI: 10.1093/gbe/evu091.

[25] S. Snir, Y.I. Wolf, and E.V. Koonin. “Universal pacemaker of genome evolution”. In: *PLoS Comput Biol* 8 (Nov. 2012), e1002785.

[26] Sagi Snir and Matteo Pellegrini. “An epigenetic pacemaker is detected via a fast conditional expectation maximization algorithm”. In: *Epigenomics* 10.6 (2018). PMID: 29979108, pp. 695–706. DOI: 10.2217/epi-2017-0130. eprint: https://doi.org/10.2217/epi-2017-0130. URL: https://doi.org/10.2217/epi-2017-0130.

[27] Sagi Snir, Bridgett M. vonHoldt, and Matteo Pellegrini. “A Statistical Framework to Identify Deviation from Time Linearity in Epigenetic Aging”. In: *PLoS Comput Biol* 12.11 (Nov. 2016), pp. 1–15. DOI: 10.1371/journal.pcbi.1005183. URL: http://dx.doi.org/10.1371%2Fjournal.pcbi.1005183.

[28] Gilbert Strang. *Introduction to Linear Algebra, Second Edition*. Wellesley-Cambridge Press, 1993. ISBN: 0961408855.

[29] Reid F. Thompson et al. “Tissue specific dysregulation of DNA methylation in aging”. In: *Aging Cell* 9.4 (2010), pp. 506–518. DOI: 10.1111/j.1474-9726.2010.00577.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1474-9726.2010.00577.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1474-9726.2010.00577.x.

[30] Richard F Walker et al. "Epigenetic age analysis of children who seem to evade aging." eng. In: *Aging* 7.5 (May 2015), pp. 334–339. ISSN: 1945-4589 (Electronic). DOI: 10.18632/aging.100744.

[31] Richard F. Walker et al. "Epigenetic age analysis of children who seem to evade aging". In: *Aging* 7.5 (2015), pp. 334–339. ISSN: 1945-4589. DOI: 10.18632/aging.100744. URL: https://doi.org/10.18632/aging.100744.

[32] Tina Wang et al. "Epigenetic aging signatures in mice livers are slowed by dwarfism, calorie restriction and rapamycin treatment". In: *Genome Biology* 18.1 (Mar. 2017), p. 57. ISSN: 1474-760X. DOI: 10.1186/s13059-017-1186-2. URL: https://doi.org/10.1186/s13059-017-1186-2.

[33] L. Wasserman. *All of Statistics*. New York: Springer, 2004. Chap. 4.

[34] Y. I. Wolf, S. Snir, and E. V. Koonin. "Stability along with Extreme Variability in Core Genome Evolution". In: *Genome Biology and Evolution* 5.7 (2013), pp. 1393–1402.

[35] Qianhua Xu and Wei Xie. "Epigenome in Early Mammalian Development: Inheritance, Reprogramming and Establishment". In: *Trends in Cell Biology* 28.3 (2018), pp. 237–253. ISSN: 0962-8924. DOI: https://doi.org/10.1016/j.tcb.2017.10.008. URL: http://www.sciencedirect.com/science/article/pii/S096289241730199X.

CHAPTER 6

The Epigenetic Pacemaker is a more sensitive tool than penalized regression for identifying factors that impact epigenetic aging

Colin Farrell[1], Charlotte Hu[2], Kyle Pu[2], Kalsuda Lapborisuth[2], Sagi Snir[3], Matteo Pellegrini[3,4]

[1]Department of Human Genetics, University of California, Los Angeles, CA, USA;

[2]Dept. of Molecular, Cell and Developmental Biology; University of California, Los Angeles, CA 90095, USA;

[3]Dept. of Evolutionary Biology, University of Haifa, Israel;

[4]Corresponding Author, matteop@mcdb.ucla.edu

DNA methylation based chronological age prediction models, epigenetic clocks, are commonly employed to study age related biology. The error between the predicted and observed age is often interpreted as a form of biological age acceleration and many studies have measured the impact of environmental factors on epigenetic age. Epigenetic clocks are fit using approaches that minimize the error between the predicted and observed chronological age. As a result, epigenetic clocks remove potentially informative signals that could be caused by external factors such as disease states. We compare the methods used to construct epigenetic clocks to an evolutionary framework of epigenetic aging, the epigenetic pacemaker (EPM). In contrast to epigenetic clocks, the EPM minimizes error across a set of methylation sites

to find an optimal epigenetic state. We show that the EPM is more sensitive to simulated age accelerating traits. We also show that the EPM is more sensitive at detecting sex and cell type effects in a large aggregrate dataset assembled from publicly available data. Thus we find that the epigenetic pacemaker is more suited to the study of factors that impact age acceleration than traditional epigenetic clocks based on linear regression models.

6.1 Introduction

Epigenetic clocks, accurate age predictions models made using DNA methylation, are promising tools for the study of aging and age related biology. The difference between the observed and expected epigenetic age can be interpreted as a measure of biological age acceleration [20]. Age acceleration observed using the first generation of epigenetic clocks [18, 15] has been associated with a variety of health outcomes including mortality risk[34, 38], cancer risk [12], cardiovascular disease[24] and other negative health outcomes[1, 21, 22]. However, as epigenetic clocks become more accurate epigenetic age acceleration is no longer associated with mortality [54]. This has led to the development of a second generation of epigenetic clocks[30, 32, 5] fit against an integrated measure of health rather than chronological age.

Generation one and two clocks were fit following the same procedure. A penalized regression model is fit against a trait of interest, whether it is age or a measure of biological age. Given an elastic net model of form $y = \beta X$, the goal of penalized regression is to maximize the likelihood by reducing the prediction error of the model, $L(\lambda_1, \lambda_2, \beta) = |y - X\beta|^2 + |\lambda_2\beta|^2 + |\lambda_1\beta|$. In the case of epigenetic clocks, the likelihood is maximized by minimizing the difference between the observed and predicted age subject to the elastic net penalty. The error minimization assumes that the dependent variable can be modeled linearly using the underlying DNA methylation data. Methylation sites that are non-linear with respect to the signal of interest or that increase modeled error will be preferentially excluded. Consequently, biologically meaningful information may be discarded during model fitting. This problem is magnified in the case of epigenetic clocks where the relationship between methylation and time is nonlinear[41]. Rather than determining how well a signal of interest can be modeled using epigenetic information, an alternative and complementary approach to studying epigenetic aging is to model how methylation changes for a predetermined collection of sites with respect to age.

To this end, we have developed the epigenetic pacemaker (EPM) [42, 13] to model methylation changes with age. Given j individuals and i methylation sites, under the EPM an individual methylation site can be modeled as $\hat{m}_{ij} = m_i^0 + r_i s_j + \epsilon_{ij}$ where $\hat{m}_{ij}$ is the observed methylation value, m_i^0 is the initial methylation value, r_i is the rate of change, s_j is the epigenetic state, and ϵ_{ij} is a normally distributed error term. The r_i and m_i^0 are characteristic of the sites across all individuals and the epigenetic state of an individual s_j is set using information from all modeled sites. Given an input matrix $\hat{M} = [\hat{m}_{i,j}]$ the EPM utilizes a fast conditional expectation maximization algorithm to find the optimal values of m_i^0, r_i, and s_j to minimize the error between the observed and predicted methylation values across a set of sites. This is accomplished by first fitting a linear model per site using age as the initial s_j. The s_j of the modeled samples is then updated to minimize the error between the observed and predicted methylation values. This process is performed iteratively until the reduction in error is below a specified threshold or the maximum number of iterations is reached. Under the EPM, the epigenetic state has a linear relationship with the modeled methylation data, but not necessarily with respect to time. This allows for non-linear relationships between time and methylation to be modeled without prior knowledge of the underlying form. Additionally, after model training each site has characteristic m_i^0 and r_i that describe the site in relation to other modeled sites and the output epigenetic states. Importantly, sites with different functional relationships between methylation and time are modeled collectively.

In the current work, we extend the EPM model using a simulation framework to simulate methylation matrices associated with age and phenotypes that drive age acceleration. Utilizing this simulation framework we evaluate the ability of penalized regression and EPM models to detect simulated traits that have linear and nonlinear associations with age and age accelerating traits. We then validate the simulation results utilizing a large aggregate dataset compiled from publicly available data.

6.2 Results

Simulation of Trait Associated Methylation Matrix

Under the EPM the epigenetic state for individual j, S_j, can be interpreted as a form of biological age that represents a weighted sum of aging associated phenotypes $S_j = \sum_{k=1}^{n} \alpha_1 p_{1,j} + ... + \alpha_k p_{k,j}$. Under this model α_k is the weight for phenotype k and $p_{k,j}$ is the value of phenotype k. Phenotypes may contribute to increased or decreased aging respectively that when considered as a whole contribute to the overall aging rate observed for an individual.

As shown in our previous work[41], the relationship between $p_{k,j}$ and time is not necessarily linear. When simulating age associated phenotypes, each phenotype can be represented as $p_{k,j} = Age_j^{\gamma_k} q_{k,j}$, where γ_k is a phenotype specific learned parameter shared among all individuals. In the formulation of a phenotype, $q_{j,k}$ is representative of exposure. The observed phenotype is modeled as an interaction between age and an exposure of varying magnitude among individuals. This formulation is flexible as non-age dependent traits can be easily simulated by setting $\gamma_k = 0, p_{k,j} = Age_j^0 q_{k,j} = q_{k,j}$.

The objective of the EPM is to find the optimal values of m_i^0, r_i, and s_j to minimize the error between the predicted and observed methylation values across a system of methylation sites. The optimal values of s_j represent a composite metric representative of each modeled site. Individual sites can be described as a linear model where $\hat{m_{i,j}} = m_i^0 + r_i P_{i,j} + \epsilon_{i,j}$. $P_{i,j}$ is a weighted sum of phenotypes influencing the methylation status of an individual site,

$P_{i,j} = \sum_{k=1}^{n} \upsilon_1 p_{1,j} + ... + \upsilon_k p_{k,j}$. Additionally, while individuals will have varying $P_{i,j}$ values, the contribution of each phenotype, but not magnitude of the exposure, is shared among all individuals.

To assess the sensitivity of the EPM and penalized regression approaches at detecting the influence of epigenetic state associated traits we simulated a methylation matrix containing linear and nonlinear age associated traits of form $p_{k,j} = Age_j^{\mathcal{N}(0.5,0.01)} q_{k,j}$ and $p_{k,j} = Age_j^{\mathcal{N}(1,0.01)} q_{k,j}$. The trait γ parameter was generated by sampling from a normal distribution $\mathcal{N}(0.5, 0.01)$ to generate traits with varying relationships with time (Figure 1). Simulated traits included binary phenotype ($P = 0.5$), continuous phenotypes influenced by age and sample health, or a continuous trait influenced by age only (Table 1). Samples were simulated by assigning an age from a uniform distribution, $\mathcal{U}(0, 100)$ and sample health from a normal distribution. The effect, q, of carrying a binary trait was varied from 0.995 to 1.0 over 5 equally spaced intervals. Given a binary trait form of $p_{k,j} = Age_j^{0.5} q_{k,j}$ a 0.001 decrease in q corresponds to a 1 percent decrease in epigenetic state by age 100 relative samples not assigned the binary trait. Within each interval the standard deviation of the sample health sampling distribution was varied from 0.0 to 0.01 over 5 equally spaced intervals. The simulation was repeated 50 times for each binary, continuous trait combination with 500 simulated samples. Additionally, at a binary q of 0.995 the range of continuous traits was expanded over a broader range to assess the model sensitivity at detecting the continuous trait. Five methylation sites for all continuous traits were then simulated and 50 methylation sites for the binary trait. Additionally, 50 sites were simulated that were equally influenced by a mixture of four continuous traits and the simulated binary trait. The resulting simulation matrix contains 450 methylation sites and 500 samples per simulation.

Given a simulation dataset the samples were split randomly and 50% of data used for training and testing respectively. EPM and penalized regression models were fit to each simulation training matrix to predict the epigenetic state and age respectively. To assess the sensitivity of each model to detect a trait influencing the epigenetic age or state with the testing data we fit a regression model where the epigenetic age or state is dependent on the age, square-root of the age, the health status, and binary trait status of the sample ($S_j = Age + \sqrt{Age} + health_j + binary_j$). The square-root of the age is included in the regression model to account for the non-linear relationship between the simulated age and methylation data.

As the exposure size of the binary trait is decreased from 1.00 to 0.995 the ability of the EPM and regression models to detect an influence on epigenetic state is improved (Figure 2 A and B). At an effect size of 0.995 the estimated effect of the binary trait is significant ($\mu = 0.041, \sigma = 0.121$) while in the penalized regression model it is not ($\mu = 0.271, \sigma = 0.278$). At an exposure size of 1.0, equivalent to no effect of the simulated binary trait, the distribution for the EPM and linear models is randomly dispersed. The binary trait model coefficients for the EPM are proportional to the effect size and the distributions are clearly defined given an increase in simulated exposure size (S.Figure 2 A). By contrast, the assigned model coefficients for the epigenetic age models are not clearly defined by exposure size (S.Figure 2 B.).

The ability to observe the health effect through the simulated continuous traits improves in both the linear and EPM models as the standard deviation of the sample health sampling distribution is increased (Figure 2 A and B). At an exposure size of 0.002 and 0.0025 the average EPM model is significant ($\mu = 0.0269, \sigma = 0.0875$) while the average linear model is not ($\mu = 0.108, \sigma = 0.209$). At a continuous trait standard deviation above 0.005 both models produce significant results. Additionally, EPM and penalized regression models were generated against a matrix of 60 sites, where ten of the sites are influenced by the binary trait ($q = 0.995$) and 50 sites were randomly selected that were influenced by sample health or only sample age. The proportion of age only to health associated sites was varied from 1 to 0.5. As the proportion of age-only sites is decreased the ability to observe the continuous trait improves for both the EPM and penalized regression models (Figure S. 2 C and D) but the ability to observe the effect is dependent on the continuous standard deviation. The EPM shows improved performance at detecting the continuous traits at the same distribution given a continuous trait standard deviation relative to the penalized regression model (Figure S. 2 C and D). Additionally, as expected the penalized regression models predominantly selected sites associated with linear traits, $p_{k,j} = Age_j^{\mathcal{N}(1,0.01)} q_{k,j}$, (S. Figure 2 A and B)

Universal Blood EPM and Penalized Regression Models

To compare the performance of the EPM and linear models in real data we assembled a large aggregate dataset of Illumina 450k array data[33, 48, 44, 27, 50, 43, 9, 19, 28, 53, 6, 11, 46] deposited in the Gene Expression Omnibus[3] (GEO). The combined datasets represented 6,251 whole blood tissue samples across 16 GEO series. We trained EPM and penalized regression models using a subset of the total aggregate data assembled from four GEO series[26, 31, 7, 10] ($n = 1605$) with samples spanning a wide age range (0.01 - 94.0 years). The training set was split by predicted sex and then was stratified and split by age, 80% of the samples from each sex were combined ($n = 1283$) for model training and the remaining samples ($n = 322$) for model evaluation. Methylation values for all samples were quantile normalized by probe type[18] using the median site methylation values across all training samples for each methylation site.

We then fit a penalized regression model to the training matrix as follows. The normalized training methylation matrix was first filtered to remove sites with a variance below 0.001, resulting in a training matrix with 183,114 sites down from the 485,512 sites in the unfiltered matrix. A cross validated elastic net model was trained against training sample ages using the filtered methylation matrix. The trained model performed well on the training and testing datasets (S.Figure 3).

In contrast to penalized regression based approaches, site selection for the EPM model is performed outside of model fitting. Methylation sites were selected for model training as follows. Sites were initially selected for modeling if the absolute Pearson correlation coefficient between methylation values and age was greater than 0.45 ($n = 9937$). A per site regression model was fit using the observed methylation value as the independent variable and age as the explanatory variable. Sites with a mean absolute error (MAE) less than 0.025 between the predicted and observed methylation values were retained for further analysis ($n = 3832$). An EPM model was then fit to using these sites (Figure 3A). We then sought to identify subsets of sites that had functionally similar forms between age and methylation, that were represented in a sufficient number of sites to model. This was done to filter sites that were associated with age by chance and to select cluster with low with low prediction error. Subsets of sites with similar functional form were clustered[14] by the euclidean distance between the single site regression model residuals using affinity propagation [14]). Cross validated EPM and penalized regression models were trained for all clusters with greater than ten sites ($n = 34$). The cluster EPM models show varying association between the epigenetic state and age relative to the EPM model fit with all sites used during clustering (Figure 3B). This resembles the simulated methylation matrices where sites with differing functional forms are modeled collectively. Clusters with an observed EPM and penalized regression MAE less than 6 ($n = 3$) were combined to fit final EPM and penalized regression models. The combined cluster EPM and combined cluster regression model performed well on the training and testing datasets (S.Fig 2). Additionally, principal component analysis (PCA) was performed on the cell type abundance estimates made for the training data. The trained PCA model was used to predict the cell type PCs for the testing and validation datasets.

We evaluated the combined cluster EPM, combined cluster penalized regression, and the full penalized regression models against a validation data set comprising 14 GEO series experiments representing 4,600 whole blood tissue samples. Each model accurately predicted the epigenetic state or epigenetic age of the validation samples (Figure 4). We then fit a ordinary least squares regression model for every validation experiment individually to predict the observed epigenetic age or state using the sample age, the square root of age, cell type PCs, and predicted sex ($S_j = Age + \sqrt{Age} + PC1 + PC2 + PC3 + Sex + Intercept$).

If the proportion of female samples to the total number of sample was greater than 0.7 the sex term was dropped from the experimental regression model. Significant cell type PC2 coefficients were observed for all EPM models and the majority of the cluster and full penalized regression models (Figure 5A). Significant cell type PC1 and PC3 coefficients were observed for the majority of the EPM models but not for the cluster or full penalized regression models. Significant sex effects ($p < 0.0038$) were observed for 9, 4, 0 out 15 models for the EPM, cluster penalized regression, and full penalized regression respectively (Figure 5B).

6.3 Discussion

The epigenetic state of an individual is dynamic, with methylation sites dispersed throughout the genome influenced in context specific ways. This is supported by the array of epigenetic clocks, fit to age and aggregate measures of health, and to the emergence of other DNA methylation based biomarkers[36, 52, 16]. Given a large enough matrix, penalized regression will select sites that minimize the prediction error given a modeled trait. Utilizing this approach will generate a model where sites that are strongly associated with the modeled trait are preferentially selected which can remove biological signals from other factors that may influence methylation. Depending on the context this approach is both more and less effective than the EPM. The penalized regression models provide more accurate age predictions ($R^2 = 0.875, 0.911$) than the EPM model ($R^2 = 0.821$), and the model output can be directly related to the age of a sample. By contrast, the predictions made by the EPM provide results as an epigenetic state that can't be directly interpreted as an age.

However, while the penalized regression models were more accurate for predicting age, the EPM was more sensitive to cell type and sex effects with real data and more sensitive to the simulated traits. The EPM also captures the nonlinear relationship between the modeled methylation data and time. Naturally, as a penalized regression model fit to age improves, the influence of other factors on epigenetic age predictions will be minimized. By contrast, the EPM minimizes the difference between the predicted and observed methylation values. The output epigenetic state will be influenced by the modeled methylation data, even if the modeled sites introduce prediction error. If the introduced error is biologically meaningful then the resulting model can be less accurate but more informative. This idea has been recently supported by two studies looking at epigenetic aging in marmots[39] and zebras[29]. EPM models showed an association between hibernation and slowed epigenetic aging in marmots and increased epigenetic age associated with zebra inbreeding; no such associations were observed with penalized regression epigenetic age models.

Ultimately, EPM and penalized regression approaches are complementary to one another. While penalized regression approaches can generate highly accurate models to predict age, these approaches naturally reduce the influence of other factors on age related biology. By contrast, the EPM models a collection of sites and is sensitive to environmental and physiological factors that influence epigenetic aging .

6.4 Methods

Simulation

The simulation framework is implemented a as python package with numpy($\geq$v1.16.3)[17] and scikit-learn($\geq$v0.220)[37] as dependencies. A simulation run results in trait associated methylation matrix and samples with assigned traits. The amount of noise added per methylation site is also retained for downstream use. The simulation procedure is implemented as follows:

1. Traits are initialized that contain the information about the trait relationship with age and a simulated sample phenotype. Given the structure $p_{k,j} = Age_j^{\gamma_k} q_{k,j}$, and k samples and j traits γ is characteristic of the trait. When a sample is passed to a

trait a value of q is generated for the sample by sampling from a normal distribution with a variance characteristic of the simulation trait. Additionally, each trait can be optionally influenced by a characteristic measure of health, h_j, for the sample. Given, a normally distributed trait $\mathcal{N}(\mu, \sigma^2)$ and a health effect h_j, the sampled distribution for individual j is $\mathcal{N}(\mu + h_j, \sigma^2)$. Continuous and binary traits can be simulated. If a binary trait is simulated a q other than 1 is assigned at a specified probability.

2. Samples are simulated by setting the age by sampling from a uniform distribution over a specified range and by setting a sample health metric h by sampling from a normal distribution centered on zero with a specified variance. Traits passed to a sample simulation object are then set according to the age and health of the sample. Simulated samples retain all the set phenotype information for downstream reference.

3. Methylation sites are simulated by randomly setting the initial methylation value, maximum observable methylation value, the rate of change at the site, and the error observed at each site. Sites are then assigned traits that influence the methylation values at each site.

4. Methylation values are simulated for each site for every individual given the simulated phenotypes with a specified amount of random noise

Simulation EPM and Penalized Regression Models

Simulation data was randomly split in half into training and testing sets. EPM models were fit using the simulated methylation matrix against age. Penalized regression models were fit using scikit-learn ElasticNet (alpha=1, l1_ratio=0.75, and selection=random). All other parameters were set to their default values. Ordinary least squares regression as implemented in statsmodels (0.11.1)[40] was utilized describe the epigenetic state or age with the following form ($S_j = Age + \sqrt{Age} + health_j + binary_j$). Full analysis is found in the EPMSimulation.ipynb supplementary file.

Methylation Array Processing

Metadata for Illumina methylation 450K Beadchip methylation array experiments deposited in the Gene Expression Omnibus (GEO) [3] with more than 50 samples were parsed using a custom python toolset. Experiments that were missing methylation beadchip array intensity data (IDAT) files, made repeated measurements of the same samples, utilized cultured cells, or assayed cancerous tissues were excluded from further processing. IDAT files were processed using minfi[2] (v1.34.0). Sample IDAT files were processed in batches according to GEO series and Beadchip identification. Methylation values within each batch were normal-exponential normalized using out-of-band probes[45]. Blood cell types counts were estimated using a regression calibration approach[23] and sex predictions were made using the median intensity measurements of the X and Y chromosomes as implemented in minfi. Whole blood array samples were used for downstream analysis if the sample median methylation probe intensity was greater than 10.5 and the difference between the observed and expected median unmethylation probe intensity is less than 0.4 given the observed median methylation intensity where the expected unmethylated signal as described by ($y = 0.66x + 3.718$).

Universal Blood EPM and Penalized Regression Models

Methylation sites with a Pearson correlation coefficient against age greater than 0.45 ($n = 9937$). A linear model was generated using numpy polyfit with age and the independent variable and methylation values as the dependent variable. Mean absolute error (MAE) was calculated as the mean absolute difference between the observed and predicted meth values according to the site linear models. A vector of residuals generated using this model was utilized for clustering by affinity propagation using scikit-learn AffitinityPropogation using a random state of 1 and a cluster preference of -2.5. All other parameters were set to default. Cross validated ($n = 10$) EPM models were generated for all training subsets.

Penalized regression models were fit using scikit-learn ElasticNetCV (cv=5 alpha=1,

*

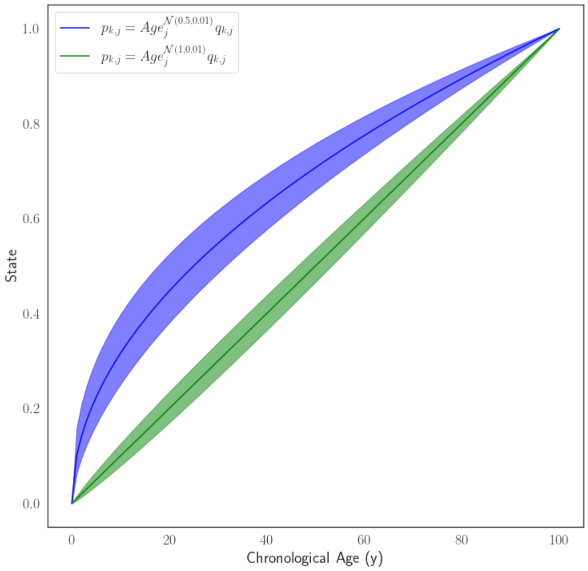
Table 1: Simulated Bisulfite Sequencing Library Parameters:
The parameters used to simulate libraries using BSBolt for tool comparisons. All simulations were carried out at read lengths of 50, 100, and 150 base pairs.

Trait Type	γ	**Health Effect**	**Age Only**	**Generated Phenotypes**
Continuous	$\mathcal{N}(0.5, 0.01)$	Yes	No	10
Continuous	$\mathcal{N}(1.0, 0.01)$	Yes	No	10
Continuous	$\mathcal{N}(0.5, 0.01)$	No	Yes	20
Continuous	$\mathcal{N}(1.0, 0.01)$	No	Yes	20
Binary	0.5	Yes	No	1

l1_ratio=0.75, and selection=random). All other parameters were set to their default values. Principal Component Analysis as implemented in scikit-learn was utilized with default parameters to perform PCA on training sample sample cell type abundances. The trained PCA was utilized to calculate cell type PCs for the testing and validation samples. Ordinary least squares regression as implemented in statsmodels (0.11.1)[40] was utilized describe the epigenetic state or age with the following form ($S_j = Age + \sqrt{Age} + CellTypePC1 + CellTypePC2 + CellTypePC3 + Sex + Intercept$). Full analysis is found in the EPMUniversalClock.ipynb supplementary file.

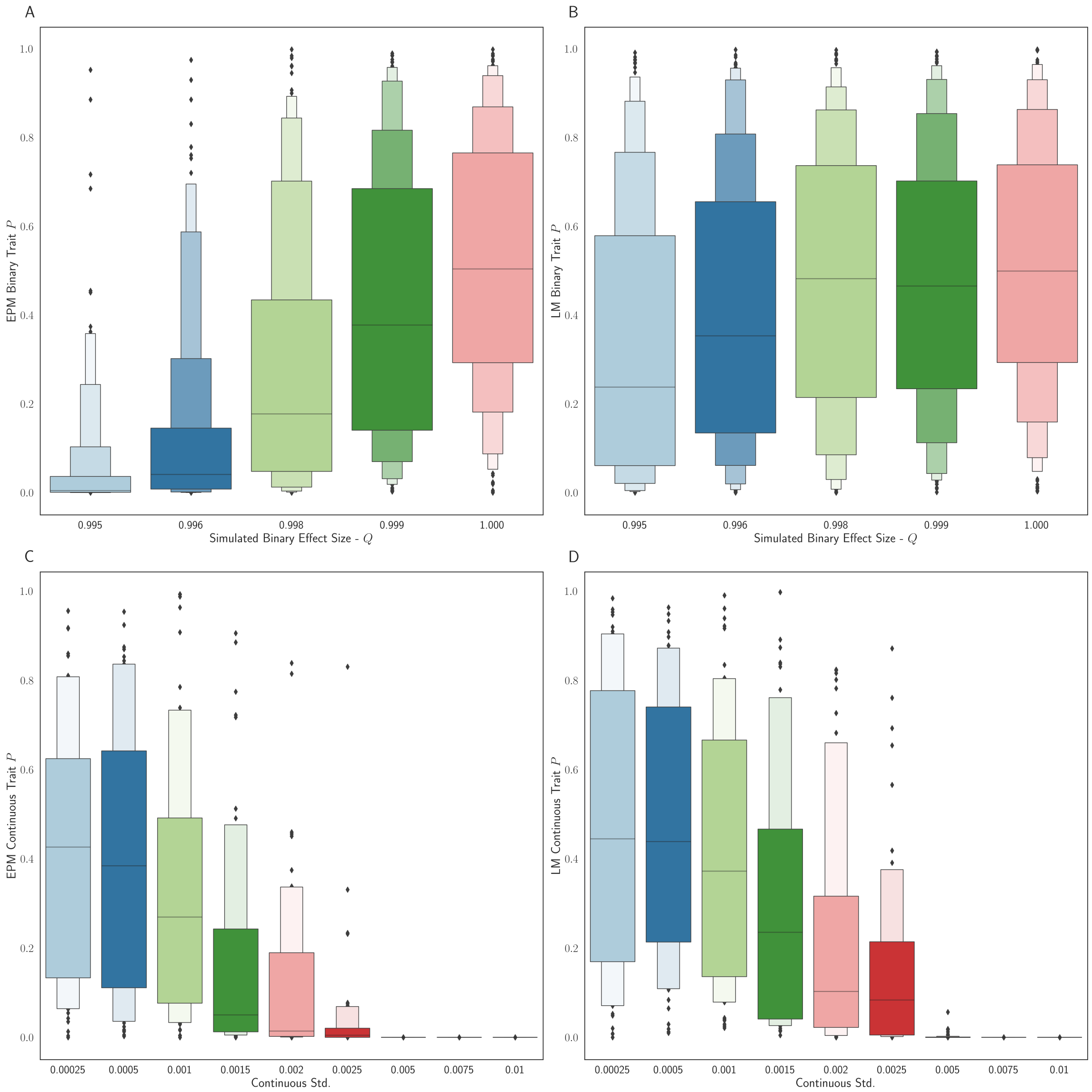
Analysis Environment

Analysis was carried out in a Jupyter[4] analysis environment. Joblib[47], SciPy[49], Matplotlib[25], Seaborn[51], Pandas[35] and TQDM[8] packages were utilized during analysis.



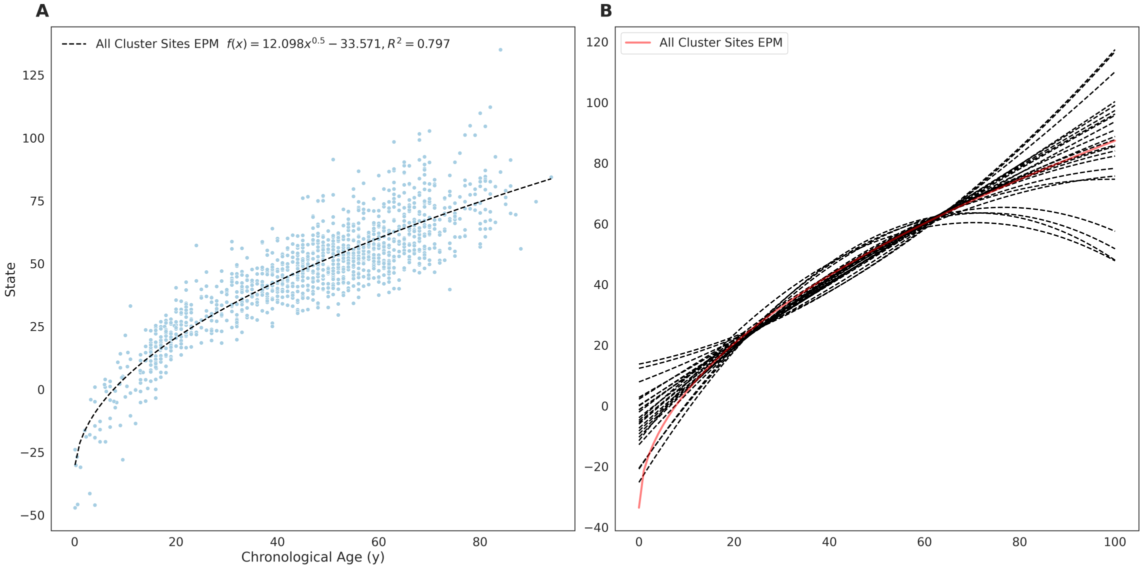
*

Figure 1: Simulated trait forms where the shaded area represent one standard deviation away from the mean γ, given $p_{k,j} = Age_j^{\gamma_k} q_{k,j}$.



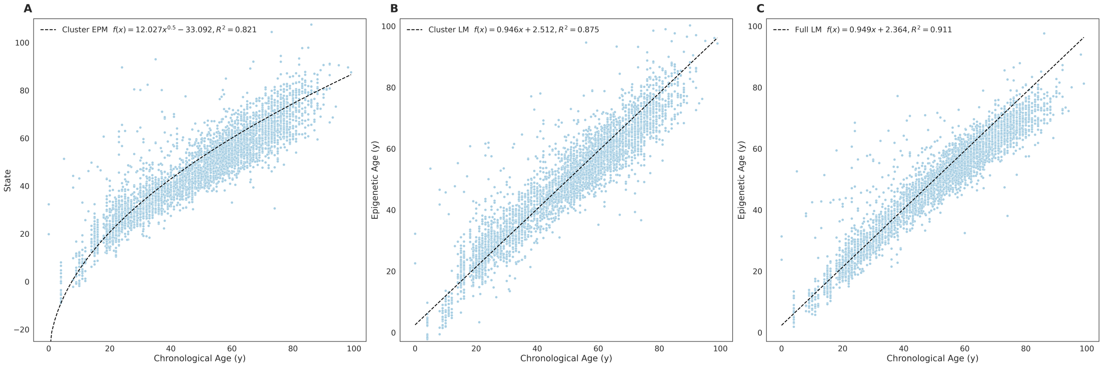
*

Figure 2: The distribution binary coefficient p-values for (A) EPM and (B) penalized regression models. The distribution of p-values given a simulation health standard deviation for (C) EPM and (D) penalized regression models



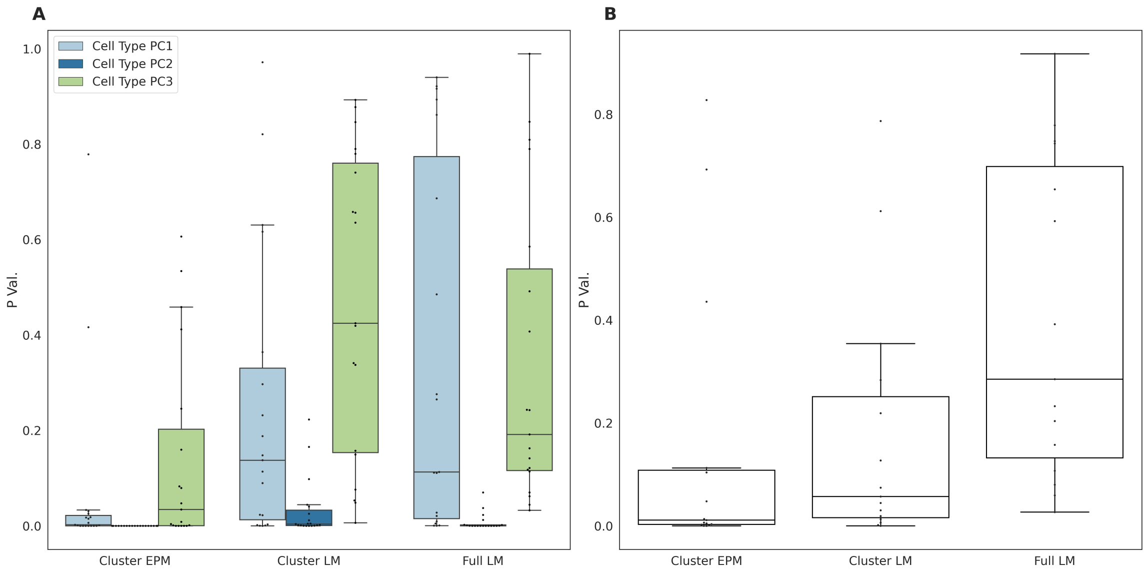
*

Figure 3: (A) EPM model fit with 3832 methylation sites with a MAE below 0.025. (C) The fit trend line for EPM clusters with more than 10 sites and an $R^2 \geq 0.4$.



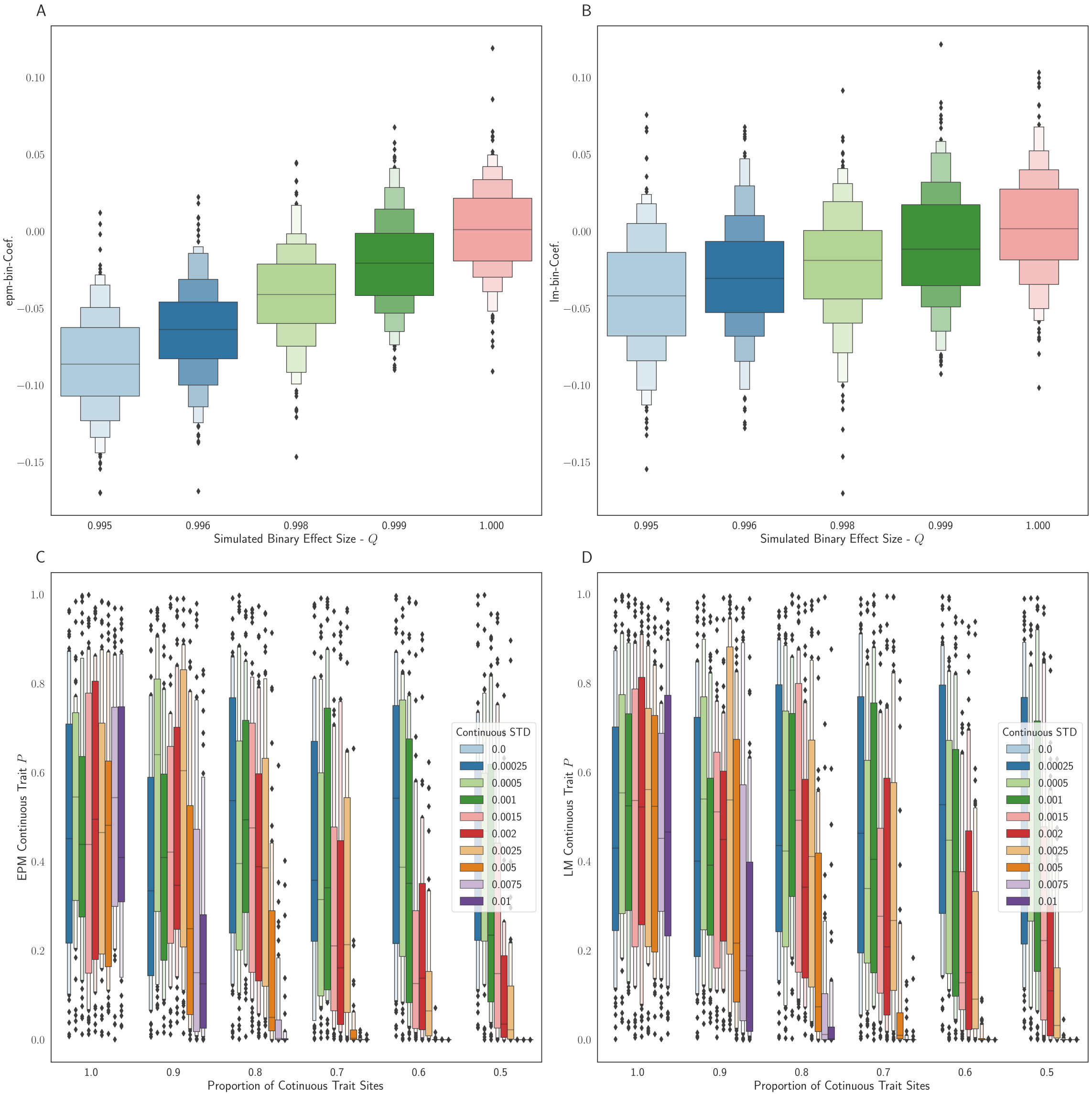
*

Figure 4: Whole blood tissue validation (A) EPM, (B) cluster penalized regression and (C) full penalized regression models.



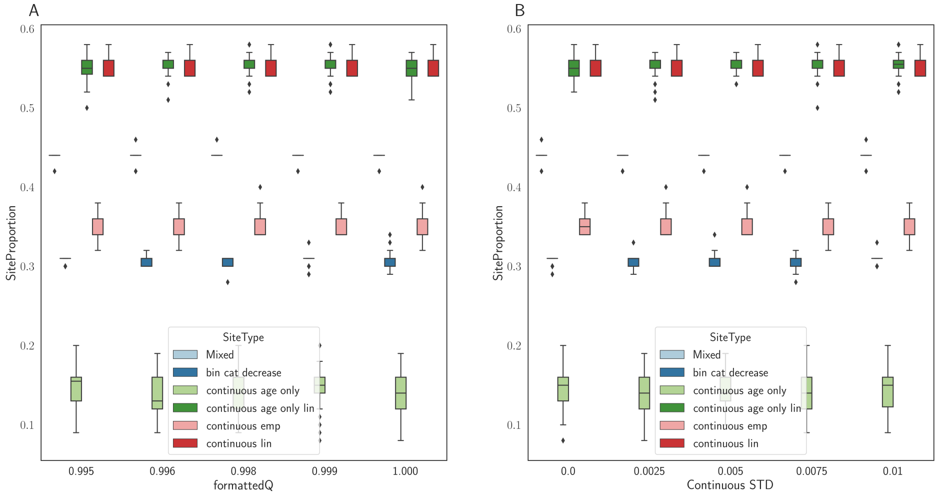
*

Figure 5: (A) Cell type principal component and (B) predicted sex regression coefficient p-values.



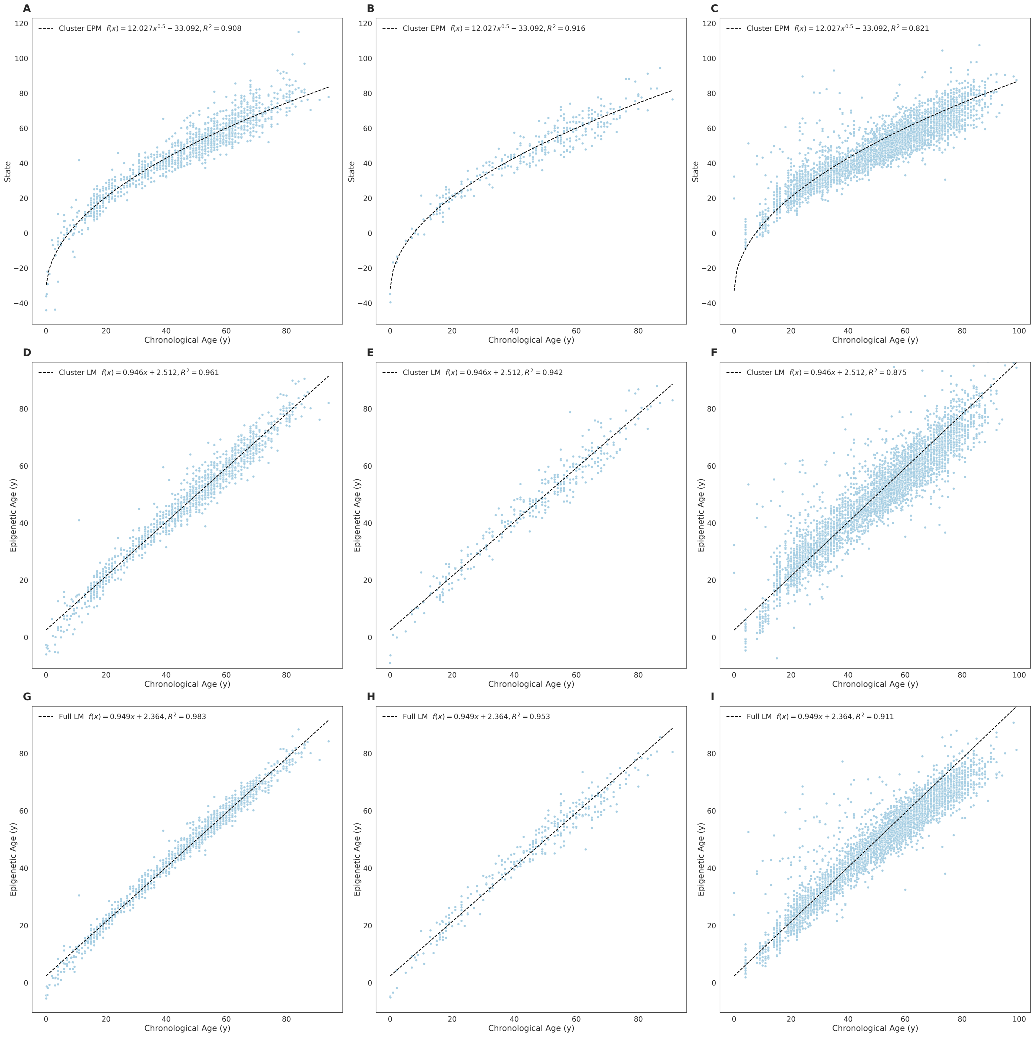
*

S. Figure 1: Binary trait coefficients in epigenetic state regression Models for EPM (A) and penalized regression (B) by binary exposure size q. Continuous health associated traits were mixed with non-health associated continuous traits at set proportions. As the proportion of health associated sites increases relative to the non-health associated sites the (C) EPM and (D) penalized regression models become more sensitive. The ability of both models to detect the simulated effect is dependent on the magnitude of the simulation effect q.



*

S. Figure 2: Proportion of sites, by type, selected for the simulated penalized regression models by (A) binary trait q and (B) sampling health standard deviation



*

S. Figure 2: (A - C) Train, testing, and validation EPM model. (D-E) Train, testing, and validation cluster penalized regression model. (G-J) Train, testing, and validation full penalized regression model.

Bibliography

[1] Nicola J Armstrong et al. “Aging, exceptional longevity and comparisons of the Hannum and Horvath epigenetic clocks”. en. In: *Epigenomics* 9.5 (May 2017), pp. 689–700.

[2] Martin J Aryee et al. “Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays”. en. In: *Bioinformatics* 30.10 (May 2014), pp. 1363–1369.

[3] Tanya Barrett et al. “NCBI GEO: archive for functional genomics data sets—update”. en. In: *Nucleic Acids Res.* 41.D1 (Nov. 2012), pp. D991–D995.

[4] Arindam Basu. *Reproducible research with jupyter notebooks.*

[5] Daniel W Belsky et al. “Quantification of the pace of biological aging in humans through a blood test, the DunedinPoAm DNA methylation algorithm”. en. In: *Elife* 9 (May 2020).

[6] Patricia R Braun et al. “Genome-wide DNA methylation comparison between live human brain and peripheral tissues within individuals”. en. In: *Transl. Psychiatry* 9.1 (Jan. 2019), p. 47.

[7] Darci T Butcher et al. “CHARGE and Kabuki Syndromes: Gene-Specific DNA Methylation Signatures Identify Epigenetic Mechanisms Linking These Clinically Overlapping Conditions”. en. In: *Am. J. Hum. Genet.* 100.5 (May 2017), pp. 773–788.

[8] Casper O da Costa-Luis. “tqdm: A Fast, Extensible Progress Meter for Python and CLI”. In: *JOSS* 4.37 (May 2019), p. 1277.

[9] Luke Dabin et al. *Altered DNA methylation profiles in blood from patients with sporadic Creutzfeldt-Jakob disease.*

[10] Estela Dámaso et al. “Comprehensive Constitutional Genetic and Epigenetic Characterization of Lynch-Like Individuals”. en. In: *Cancers* 12.7 (July 2020).

[11] Christiana A Demetriou et al. “Methylome analysis and epigenetic changes associated with menarcheal age”. en. In: *PLoS One* 8.11 (Nov. 2013), e79391.

[12] Pierre-Antoine Dugué et al. “DNA methylation-based biological aging and cancer risk and survival: Pooled analysis of seven prospective studies”. en. In: *Int. J. Cancer* 142.8 (Apr. 2018), pp. 1611–1619.

[13] Colin Farrell, Sagi Snir, and Matteo Pellegrini. “The Epigenetic Pacemaker: modeling epigenetic states under an evolutionary framework”. en. In: *Bioinformatics* 36.17 (Nov. 2020), pp. 4662–4663.

[14] Brendan J Frey and Delbert Dueck. “Clustering by passing messages between data points”. en. In: *Science* 315.5814 (Feb. 2007), pp. 972–976.

[15] Gregory Hannum et al. “Genome-wide methylation profiles reveal quantitative views of human aging rates”. en. In: *Mol. Cell* 49.2 (Jan. 2013), pp. 359–367.

[16] Xiaoke Hao et al. “DNA methylation markers for diagnosis and prognosis of common cancers”. en. In: *Proc. Natl. Acad. Sci. U. S. A.* 114.28 (July 2017), pp. 7414–7419.

[17] Charles R Harris et al. “Array programming with NumPy”. en. In: *Nature* 585.7825 (Sept. 2020), pp. 357–362.

[18] Steve Horvath. *DNA methylation age of human tissues and cell types.* 2013.

[19] Steve Horvath and Andrew J Levine. “HIV-1 Infection Accelerates Age According to the Epigenetic Clock”. en. In: *J. Infect. Dis.* 212.10 (Nov. 2015), pp. 1563–1573.

[20] Steve Horvath and Kenneth Raj. “DNA methylation-based biomarkers and the epigenetic clock theory of ageing”. en. In: *Nat. Rev. Genet.* 19.6 (June 2018), pp. 371–384.

[21] Steve Horvath et al. “Decreased epigenetic age of PBMCs from Italian semi-supercentenarians and their offspring”. en. In: *Aging* 7.12 (Dec. 2015), pp. 1159–1170.

[22] Steve Horvath et al. “Obesity accelerates epigenetic aging of human liver”. en. In: *Proc. Natl. Acad. Sci. U. S. A.* 111.43 (Oct. 2014), pp. 15538–15543.

[23] Eugene Andres Houseman et al. “DNA methylation arrays as surrogate measures of cell mixture distribution”. en. In: *BMC Bioinformatics* 13 (May 2012), p. 86.

[24] Rae-Chi Huang et al. “Epigenetic Age Acceleration in Adolescence Associates With BMI, Inflammation, and Risk Score for Middle Age Cardiovascular Disease”. en. In: *J. Clin. Endocrinol. Metab.* 104.7 (July 2019), pp. 3012–3024.

[25] John D Hunter. *Matplotlib: A 2D Graphics Environment.* 2007.

[26] Asa Johansson, Stefan Enroth, and Ulf Gyllensten. “Continuous Aging of the Human DNA Methylome Throughout the Human Lifespan”. en. In: *PLoS One* 8.6 (June 2013), e67378.

[27] Randi K Johnson et al. “Longitudinal DNA methylation differences precede type 1 diabetes”. en. In: *Sci. Rep.* 10.1 (Feb. 2020), p. 3721.

[28] Yuko Kurushima et al. *Epigenetic findings in periodontitis in UK twins: a cross-sectional study.* 2019.

[29] Brenda Larison et al. “Epigenetic models predict age and aging in plains zebras and other equids”. en. Mar. 2021.

[30] Morgan E Levine et al. “An epigenetic biomarker of aging for lifespan and healthspan”. en. In: *Aging* 10.4 (Apr. 2018), pp. 573–591.

[31] Yun Liu et al. *Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis.* 2013.

[32] Ake T Lu et al. “DNA methylation GrimAge strongly predicts lifespan and healthspan”. en. In: *Aging* 11.2 (Jan. 2019), pp. 303–327.

[33] Francesco Marabita et al. “Author Correction: Smoking induces DNA methylation changes in Multiple Sclerosis patients with exposure-response relationship”. en. In: *Sci. Rep.* 8.1 (Mar. 2018), p. 4340.

[34] Riccardo E Marioni et al. “DNA methylation age of blood predicts all-cause mortality in later life”. en. In: *Genome Biol.* 16 (Jan. 2015), p. 25.

[35] Wes McKinney. *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*. en. “O’Reilly Media, Inc.”, Oct. 2012.

[36] Luz D Orozco et al. *Epigenome-wide association in adipose tissue from the METSIM cohort*. 2018.

[37] F Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *J. Mach. Learn. Res.* 12 (2011), pp. 2825–2830.

[38] Laura Perna et al. *Epigenetic age acceleration predicts cancer, cardiovascular, and all-cause mortality in a German case cohort*. 2016.

[39] Gabriela M Pinho et al. “Hibernation slows epigenetic aging in yellow-bellied marmots”. en. Mar. 2021.

[40] Skipper Seabold and Josef Perktold. “Statsmodels: Econometric and statistical modeling with python”. In: *Proceedings of the 9th Python in Science Conference*. Vol. 57. 2010, p. 61.

[41] Sagi Snir, Colin Farrell, and Matteo Pellegrini. “Human epigenetic ageing is logarithmic with time across the entire lifespan”. en. In: *Epigenetics* 14.9 (Sept. 2019), pp. 912–926.

[42] Sagi Snir, Bridgett M vonHoldt, and Matteo Pellegrini. “A Statistical Framework to Identify Deviation from Time Linearity in Epigenetic Aging”. en. In: *PLoS Comput. Biol.* 12.11 (Nov. 2016), e1005183.

[43] Carolina Soriano-Tárraga et al. “Epigenome-wide association study identifies TXNIP gene associated with type 2 diabetes mellitus and sustained hyperglycemia”. en. In: *Hum. Mol. Genet.* 25.3 (Feb. 2016), pp. 609–619.

[44] Qihua Tan et al. “Epigenetic signature of birth weight discordance in adult twins”. en. In: *BMC Genomics* 15 (Dec. 2014), p. 1062.

[45] Timothy J Triche Jr et al. “Low-level processing of Illumina Infinium DNA Methylation BeadArrays”. en. In: *Nucleic Acids Res.* 41.7 (Apr. 2013), e90.

[46] Liina Tserel et al. “Age-related profiling of DNA methylation in CD8+ T cells reveals changes in immune response and transcriptional regulator genes”. en. In: *Sci. Rep.* 5 (Aug. 2015), p. 13107.

[47] Gaël Varoquaux and O Grisel. “Joblib: running python function as pipeline jobs”. In: *packages. python. org/joblib* (2009).

[48] N T Ventham et al. “Integrative epigenome-wide analysis demonstrates that DNA methylation may mediate genetic risk in inflammatory bowel disease”. en. In: *Nat. Commun.* 7 (Nov. 2016), p. 13507.

[49] Pauli Virtanen et al. “SciPy 1.0: fundamental algorithms for scientific computing in Python”. In: *Nat. Methods* (Feb. 2020).

[50] Sarah Voisin et al. *Many obesity-associated SNPs strongly associate with DNA methylation changes at proximal promoters and enhancers.* 2015.

[51] Michael Waskom. “seaborn: statistical data visualization”. In: *J. Open Source Softw.* 6.60 (Apr. 2021), p. 3021.

[52] Tarryn Willmer et al. “Corrigendum: Blood-Based DNA Methylation Biomarkers for Type 2 Diabetes: Potential for Clinical Applications”. en. In: *Front. Endocrinol.* 10 (Jan. 2019), p. 1.

[53] Anthony S Zannas et al. “Epigenetic upregulation of FKBP5 by aging and stress contributes to NF-κB–driven inflammation and cardiovascular risk”. en. In: *Proc. Natl. Acad. Sci. U. S. A.* 116.23 (June 2019), pp. 11370–11379.

[54] Qian Zhang et al. *Improved precision of epigenetic clock estimates across tissues and its implication for biological ageing.* 2019.