ENVIRONMENTAL-RISK (E-RISK) LONGITUDINAL TWIN STUDY
CONCEPT PAPER FORM

Proposing Author: Karen Sugden

Author's affiliation, phone, and e-mail address:

Department of Psychology & Neuroscience, Duke University

Sponsoring Investigator (if the proposing author is a student, a post-doc or a colleague):

Avshalom Caspi

Proposed co-authors:

Jonathan Mill, Eilis Hannon,
David Corcoran, Joey Prinz, Ben Williams, Dan Belsky, Terrie Moffitt, Line Rasmussen
Chloe Wong, Helen Fisher, Louise Arseneault
Richie Poulton

Provisional Paper Title: Assessing the scale and impact of differential reliability of DNA methylation levels measured using Illumina BeadChips (Very provisional. We promise to come us with something more enticing!).

Date: January 10, 2019

Objective of the study and its significance:

Measurement reliability is of critical importance to the reproducibility of research, and a core component of ensuring high reliability is to standardize protocols. However, a barrier to the standardization of repeated measures, be it between research sites or over time, is that protocols undergo improvement or become obsolete. One example is the commercial epigenome-wide DNA methylation arrays produced by Illumina. These arrays are the standard tool for the assessment of global DNA methylation patterns; however, they undergo relatively regular updates leading to obsolescence of products obviating the continued employment of standard procedures.
Previous research suggests that the reliability of individual-level probe measurements between iterations of Illumina DNA methylation arrays varies substantially[1]. However, the impact of this variation on research reproducibility has not been fully established. In this proposal, we will assess reliability of individual DNA methylation probe measures using data on 350 E-Risk twins measured twice; once using the Human 450K BeadChip, and again using the MethylationEPIC BeadChip. We will define the reliability of individual-level CpG probe measurements through correlation between repeat measures of probe beta values measured on the two arrays. Analysis will be restricted to the intersect of probes present on both.

The proposed work will be divided into two main aims:

Aim 1: Describing the landscape of CpG probe reliability

1.1: Reliability of individual CpG probe measurements varies across the array
This section will form the basis of the analyses that follow by documenting the array-wide distribution of reliabilities.

1.2: Probe-specific characteristics are related to reliability
In this section, we will address whether there are specific properties of probes that associate with reliability. To achieve this, we will assess the distribution of reliabilities by mean beta level and variability (Standard Deviation) of methylation probes.

1.3: Spatial distribution of reliability
This section will determine if reliability is differentially distributed across genic regions. We will plot the distribution of probe reliabilities as a function of genic region annotation (3'UTR, 5'UTR, CDS, exon, intron, TSS and intergenic). Since this analysis is potentially confounded by the uneven distribution of Type I and Type II probes across the genome[2], compounded by differential reliability of the two probe types, we will further document regional distribution of probe reliabilities as a function of probe type.

Aim 2: Evaluating the impact of variation in reliability.

2.1: How does reliability influence the ability to detect genetic and environmental effects on the epigenome?
This section will incorporate the information we have about the heritability of each probe in E-Risk. Using data from ACE models[3], we will test the hypothesis that probes with low reliabilities will associate with E more than probes with high reliabilities (since in ACE models the E term incorporates random error and unreliable probes are more likely to have large error). In addition, we will also test the hypothesis that more reliable probes may be more likely to be under genetic influence (high A) and that detection of heritability is limited by unreliability.

2.2.: How does reliability affect the ability to detect developmental changes in DNA methylation?
This section will outline how probe reliability associates with change in DNA methylation over time, and we will test the hypothesis that detection of true change is limited by unreliability. We will document our ability to detect true change over and above error through probe-level analysis of the correlation of age 26 and 38 DNA methylation values in the Dunedin Study. We will analyze intra-individual change (i.e. methylation at ages 26-38) as a function of reliability (determined by analysis of the E-Risk Study, above).

2.3: What implications does probe reliability have for association testing?
In this section, we will test the hypothesis that association testing of DNA methylation is hindered by inclusion of probes with low reliabilities. To achieve this, we will turn to the composite smoking methylation score (SmPEGS)[4] in the E-Risk data. We will calculate SmPEGS in two ways; first we will use the 'native' score naive to reliability status of the 2,623 constituent probes. Second, we will generate a 'reliable' score, including only probes with reliabilities greater than 0.7. We will assess the performance of each score to discriminate smokers and non-smokers using AUC analysis, with the hypothesis that the 'reliable' score will perform better than the 'native' score.

2.4: What is the relationship between reliability of DNA methylation probes and gene expression?
This section looks at the relationship between DNA methylation probes and gene expression probesets in the Dunedin Study data. We will test the hypothesis that DNA methylation probes with higher reliability are more likely to index meaningful variation in gene expression. To achieve this, we will tabulate the correlation between gene expression probeset values of all genes and DNA methylation beta values of every CpG probe in *cis* to that gene, as a factor of methylation probe reliability (as determined using E-Risk Study data). We hypothesize that highly correlated methylation probe-gene expression probeset pairs will be over-represented by high reliability DNA methylation probes. To refine this analysis, we will annotate the associations in respect to genes known to be expressed in blood. We expect that highly correlated methylation probe-gene expression probeset pairs that represent genes known to be expressed in blood to be further over-represented by high reliability DNA methylation probes.

2.5: Are publically available DNA methylation algorithms more likely to be comprised of reliable probes?
For this section, we will analyze the distribution of reliability of probes that constitute four established DNA methylation clocks: a) the aging clock proposed by Hannum et al (2013)[5], b) the DNAmAge clock proposed by Horvath (2013)[6], c) the Biological Aging clock proposed by Levine et al.(2018)[7], and d) the aging clock proposed by Bell et al (2012)[8]. We will test the hypothesis that these algorithms preferentially select reliable probes during development, since in order to reliably index the phenotype of interest (for example, aging), the underlying probe values must be reliably measured.

2.6: Do unreliable probes mask the ability to identify DMRs?
For this section, we will test the hypothesis that probes with low reliability interfere with the correlatory relationship of adjacent CpGs that would contribute to the identification of DMRs. We will show that by restricting datasets to probes with high reliability, one is able to identify more DMRs than would be possible by including data from all available probes.

Statistical analyses:

DNA methylation dataset creation: Pre-normalized E-Risk 450K DNA methylation data will first be subset to the 350 individuals that constitute the EPICBeadChip dataset. We will then create two datasets; one in which the two array-level datasets are normalized separately and one where they are normalized together. We will employ the datasets that were normalized separately in the analyses (to 'mimic' what most researchers would encounter in terms of multiple DNA methylation datasets), but also document the metrics for data normalized together.

Variables Needed at Which Ages (names and labels):

Study: **E-Risk**

450K DNA methylation data measured in blood at age 18 (entire dataset of N=1658)
EPIC BeadChip DNA Methylation data measured in blood at age 18 (restricted dataset of N=350)
Estimated cell count proportions at age 18 (as derived from methylation data)

Results of ACE models of DNA methylation probes at age 18 (Hannon et al, 2018)

Sampsex: sex
Smkcure18: current smoking at age 18
Smkpkyre18: smoking pack years at age 18
ZresidPGS18: SmPEGs score at age 18

Dunedin Study:

DNA methylation data at Age 26
DNA methylation data at Age 38
Gene expression data at Age 38

Sex

White blood cell counts at Age 38:
Neutrophils38np
lymphocytes38np
monocytes38np
Eosinophils38np
basophils38np

References cited:

1. Logue MW, Smith AK, Wolf EJ, Maniates H, Stone A, Schichman SA *et al.* The correlation of methylation levels measured using Illumina 450K and EPIC BeadChips in blood samples. *Epigenomics* 2017; **9**(11)**:** 1363-1371.

2. Bose M, Wu C, Pankow JS, Demerath EW, Bressler J, Fornage M *et al.* Evaluation of microarray-based DNA methylation measurement using technical replicates: the Atherosclerosis Risk In Communities (ARIC) Study. *BMC Bioinformatics* 2014; **15:** 312.

3. Hannon E, Knox O, Sugden K, Burrage J, Wong CCY, Belsky DW *et al.* Characterizing genetic and environmental influences on variable DNA methylation using monozygotic and dizygotic twins. *PLoS Genet* 2018; **14**(8)**:** e1007544.

4. Sugden K, Hannon E, Arseneault L, Belsky DW, Broadbent J, Corcoran DL *et al.* Establishing a Generalized Polyepigenetic Biomarker for Tobacco Smoking. *Translational Psychiatry* 2019.

5. Hannum G, Guinney J, Zhao L, Zhang L, Hughes G, Sadda S *et al.* Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol Cell* 2013; **49**(2)**:** 359-367.

6. Horvath S. DNA methylation age of human tissues and cell types. *Genome Biol* 2013; **14**(10)**:** R115.

7. Levine ME, Lu AT, Quach A, Chen BH, Assimes TL, Bandinelli S *et al.* An epigenetic biomarker of aging for lifespan and healthspan. *Aging (Albany NY)* 2018; **10**(4)**:** 573-591.

8. Bell JT, Tsai PC, Yang TP, Pidsley R, Nisbet J, Glass D *et al.* Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population. *PLoS Genet* 2012; **8**(4)**:** e1002629.

Data Security Agreement

Provisional Paper Title	Assessing the scale and impact of differential reliability of DNA methylation levels measured using Illumina BeadChips
Proposing Author	Karen Sugden
Today's Date	January 10, 2019

Please keep one copy for your records
(Please initial your agreement)

__x__ I am familiar with the King's College London research ethics guidelines (https://www.kcl.ac.uk/innovation/research/support/ethics/about/index.aspx) and the MRC good research practice guidelines (https://www.mrc.ac.uk/research/policies-and-guidance-for-researchers/good-research-practice/).

__x__ My project has ethical approval from my institution.

___x_ I am familiar with the EU General Data Protection Regulation (https://mrc.ukri.org/documents/pdf/gdpr-guidance-note-3-consent-in-research-and-confidentiality/), and will use the data in a manner compliant with its requirements.

___x_ My computer is (a) encrypted at the hard drive level, (b) password-protected, (c) configured to lock after 15 minutes of inactivity, AND (d) has an antivirus client which is updated regularly.

____x I will treat all data as "restricted" and store in a secure fashion.

___x_ I will not share the data with anyone, including students or other collaborators not specifically listed on this concept paper.

___x_ I will not merge data from different files or sources, except where approval has been given by the PI.

___x_ I will not post data online or submit the data file to a journal for them to post.
Some journals are now requesting the data file as part of the manuscript submission process. The E-Risk Study cannot be shared because the Study Members have not given informed consent for unrestricted open access. Speak to the study PI for strategies for dealing with data sharing requests from Journals.

___x_ Before submitting my paper to a journal, I will submit my draft manuscript and scripts for data checking, and my draft manuscript for co-author mock review, allowing three weeks.

____x I will submit analysis scripts and new variable documentation to project data manager after the manuscript gets accepted for publication.

____x I will delete the data after the project is complete.

___x_ **For projects using location data:** I will ensure geographical location information, including postcodes or geographical coordinates for the E-Risk study member's homes or schools, is never combined or stored with any other E-Risk data (family or twin-level data)

___x_ **For projects using genomic data:** I will only use the SNP and/or 450K data in conjunction with the phenotypes that have been approved for use in this project at the concept paper stage.

Signature: ....k sugden, .a. caspi...