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

Proposing Author:
Jenny van Dongen

Author's affiliation, phone, and e-mail address:
Dept. of Biological Psychology, Vrije Universiteit Amsterdam, Email: j.van.dongen@vu.nl, +31-20-598 3570

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

Proposed co-authors: From From Duke University: Terrie Moffitt, Karen Sugden, Avshalom Caspi
From University of Exeter: Jon Mill, Eilis Hannon
From VU Amsterdam: Jenny van Dongen, Hamdi Mbarek, Dorret Boomsma
From Kings College London: Juan Castillo-Fernandez, Jordana Bell

Provisional Paper Title: Epigenome-wide association study of twinning

Date: 10-10-2018

Objective of the study and its significance:

Recently, a GWAS meta-analysis by the Twinning GWAS consortium has identified the first genetic variants associated with dizygotic (DZ) twinning (Mbarek et al 2016 Am J Hum Genet). Monozygotic (MZ) twins result when an embryo splits very early in development, for reasons that are currently unknown. It has been hypothesized that epigenetic mechanisms might play a role in the MZ twinning event. The role of epigenetic mechanisms in DZ twinning is unknown.

To examine the role of epigenetic mechanisms in twinning, we performed an EWAS comparing DNA methylation levels in blood between MZ and DZ twins from the Netherlands Twin Register (NTR). This analysis identified 243 sites with differential methylation level between MZ and DZ twins. We computed twin correlations to examine the within-pair resemblance of MZ and DZ twins for DNA methylation at these 243 sites, and found that for many sites, the MZ correlation is > 2 times the DZ twin correlation. Comparison with previously published putative metastable epi-alleles exhibiting 'epigenetic supersimilarity in MZ twins' (Baak et al 2018), showed that ~10% of the 243 differentially methylated sites overlap with such locations. It has been hypothesized that methylation at these sites is established prior to embryonic cleavage in MZ twins. Further overlap analyses point to enrichment in transcriptionally repressed regions and binding sites of transcription factors with a role in early embryonic development.

We like to follow-up 2 findings (methylation differences between MZ and DZ twins and correlations between MZ and DZ twins) in the E-risk Twin Study. Besides E-Risk, we have also asked our collaborators from the twinning GWAS consortium for replication, and so far TwinsUK is participating in this project too.

We are also currently carrying out additional analyses comparing methylation level of MZ and DZ twins to singletons, to examine which methylation signatures are specific to MZ twins and which are specific to DZ twins.

This study will provide novel insight into the role of epigenetic mechanisms in twinning.

Statistical analyses:

Analyses will performed at Duke University. Summary statistics (results) from these analyses will be provided to Jenny van Dongen.

1) EWAS

Sample selection

1. Twins with data available for all covariates
2. all DZ twins + **1 randomly selected MZ twin per family**

EWAS model

For all genome-wide autosomal methylation sites, please run the following model in R, with the R-package gee:

gee(CpGi~ Zygosity +BMI+age+smoking+sex+ WBC percentages + Technical + Cohort Specific, data=data, id=familynumber, family=gaussian, corstr="exchangeable", maxiter=100, na.action=na.omit,silent=TRUE)

CpGi= Methylation beta-value for CpG i

Zygosity = Zygosity, coded as 1=MZ, 0=DZ.
BMI= Body mass index
Age= Age when DNA sample was collected (not necessary to include if all samples were collected at the same age)
WBC= White blood cell percentage in the same blood sample from which DNA was extracted. If you did not measure white blood percentages in the same sample as used for the DNA methylation measurement, please estimate WBC percentages using a prediction method (e.g. Houseman's reference based method). For computational reasons, please do not include multiple WBC that are highly correlated with each other or that show very little variation between people in your cohort. For example, in the NTR, we use the following WBC as covariates: monocyte percentage, eosinophil percentage, neutrophil percentage.
Technical covariates + Cohort Specific covariates Please correct for technical (batch) covariates and other cohort-specific covariates as you deem necessary. For example, at the NTR, we include 450k array row and (bisulfite) sample plate.
Smoking= Smoking status at the moment of blood sampling, 3 levels: 0=never smoked, 1=former smoker, 2=current smoker.

2) Twin Correlations

Sample selection

1. Twins with data for all covariates
2. all complete DZ twin pairs and all complete MZ twin pairs

Twin correlations

For the 243 CpG sites, compute the correlation for methylation level between MZ and DZ twins. In this analysis, please:

- Correct for the same covariates as in the EWAS
- Provide unbounded estimates of the twin correlations (i.e. twin correlations should not be bounded between 0-1).

An R-script can be provided by Jenny to run this analysis.

Variables Needed at Which Ages (names and labels):
Age 18:
Whole blood DNA methylation (Illumina 450k array)
Zygosity
Covariates: Sex, bodymass index (BMI), smoking status, measured or predicted (Houseman) cell counts, technical (batch) covariates for the Illumina 450k array data.

Variables Needed at Which Ages (names and labels):

Study:

Age 5:

Age 7:

Age 10:

Age 12:

Age 18:

Whole blood DNA methylation (Illumina 450k array)
Zygosity
Covariates: Sex, age, bodymass index (BMI), smoking status, measured or predicted (Houseman) cell counts, technical (batch)
covariates for the Illumina 450k array data.

References cited:

Mbarek, Hamdi, et al. "Identification of common genetic variants influencing spontaneous dizygotic twinning and female fertility." The American Journal of Human Genetics 98.5 (2016): 898-908.

Van Baak, Timothy E., et al. "Epigenetic supersimilarity of monozygotic twin pairs." Genome biology 19.1 (2018): 2.

Data Security Agreement

Provisional Paper Title	Epigenome-wide association study of twinning
Proposing Author	Jenny van Dongen
Today's Date	10-10-2018

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

____ 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/).

____ My project has ethical approval from my institution.

____ 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.

____ 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.

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

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

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

____ 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.

____ 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.

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

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

____ **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)

____ **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: Jenny van Dongen.....

CONCEPT PAPER RESPONSE FORM

A. To be completed by the proposing author

Proposing Author:

x I have read the E-Risk data-sharing policy guidelines and agree to follow them

Provisional Paper Title:
Epigenome-wide association study of twinning

Potential co-authors:
Dongen, Mbarek, Boomsma, Moffitt, Sugden, Caspi, Mill, Hannon, Castillo-Fernandez, Bell

Potential Journals: to be decided

Intended Submission Date (month/year): Early 2019

Please keep one copy for your records and return one to Louise (louise.arseneault@kcl.ac.uk)

B. To be completed by potential co-authors:

☐ Approved ☐ Not Approved ☐ Let's discuss, I have concerns

Comments:

Please check your contribution(s) for authorship:

☐ Conceptualizing and designing the longitudinal study

☐ Conceptualizing and collecting one or more variables

☐ Data collection

☐ Conceptualizing and designing this specific paper project

☐ Statistical analyses

☐ Writing

☐ Reviewing manuscript drafts

☐ Final approval before submission for publication

☐ Acknowledgment only, I will not be a co-author

Signature: .....