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

Proposing Author: Matthew Suderman

Author's affiliation, phone, and e-mail address:
Integrative Epidemiology Unit, Population Health Sciences, Bristol Medical School, University of Bristol
+44 (0) 1173310090
matthew.suderman@bristol.ac.uk

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

Proposed co-authors:
Jean Golding, Marcus Pembrey, Sarah Watkins, Thomas Jonkman, Bastiaan T Heijmans, Yasmin Iles-Caven, Steven Gregory, Karen Sugden, Helen Fisher

Provisional Paper Title:
FRAXA/E repeat length, DNA methylation and neurophenotypes in non-clinical populations

Date: June 13, 2022

Note this is a request for attempted replication in E-Risk of a finding already obtained in ALSPAC. Karen Sugden has very kindly agreed to conduct the E-Risk analysis when her other commitments permit

Objective of the study and its significance:

Fragile X syndrome is a genetic disorder characterized by mental impairment that affects males more severely than females. It is caused by silencing of the FMR1 gene on the X chromosome, typically caused by having 200 or more repetitions of CGG triplet within the FRAXA region of FMR1. Multiple lines of evidence indicate that silencing is mediated by DNA methylation of the FMR1 promoter and that FMR1 expression can be rescued by reducing DNA methylation. Fragile XE syndrome is similar but less common, less severe, and typically caused by high numbers of CCG repeats in the FRAXE region at the AFF2 (FMR2) gene.

Although repeat numbers in FRAXA and FRAXE are known to vary in the general population, typically below 40 in FRAXA and below 30 in FRAXE, little is known about the extent of this variation and how it might relate to DNA methylation and physical and mental health.

To answer this question, we have measured FRAXA and FRAXE repeats in nearly 5000 males in the Avon Longitudinal Study of Parents and Children (ALSPAC). Genome-wide DNA methylation has been measured in blood samples collected from about 1179 of these males at age 15-17 using the Illumina HumanMethylation450 and EPIC Beadchips.

We observed no evidence of an association of DNA methylation with FRAXA repeats anywhere in the genome, but did observe extremely strong associations with FRAXE at two CpG sites cg20321768 and cg25587058 in the AFF2 gene promoter ($p < 1e-33$). Associations at these two sites were replicated in DNA methylation measured in blood samples collected from male ALSPAC participants at birth, age 7 and age 24 ($p < 6e-23$, n=382-472).

The functional relevance of DNA methylation at these two sites was evaluated by testing associations with

expression of the AFF2 gene in external studies: BIOS (whole blood, n=1441), MESA (monocytes, n=596) and ROSMAP (dorsolateral prefrontal cortex, n=120). Positive associations were observed in whole blood (p=0.08, p=0.003) and monocytes (effect > 0.21, p=0.02-0.07), and a negative association was observed in dorsolateral prefrontal cortex (effect < -2, p=0.1-0.22)

A potential role for FRAXE repeat length and DNA methylation at cg20321768 and cg25587058 in physical and mental health was evaluated by testing associations with 1105 phenotypes from across six domains: cognition, motor ability, personality, sources of addiction, temperament and behavior, and traits of psychiatric disorders. More specifically, associations with FRAXE repeat length were tested in one subset of ALSPAC (n=3806), and phenotypes with associations ($p < 0.01$, 24 phenotypes) were then tested for associations with DNA methylation at cg20321768 and cg25587058 measured at age 15-17 in the remaining subset of ALSPAC (n=1179). One phenotype was found to be associated with cg20321768 after adjustment for multiple tests (effect=-0.11, Bonferroni adjusted p=0.025). There was also evidence for an association with cg25587058 (effect=-0.05, Bonferroni adjusted p=0.067). The phenotype in question was "number of errors identifying high intensity angry faces" from the Diagnostic Analysis of Nonverbal Accuracy (DANVA) administered to ALSPAC participants at age 8.

We would like to determine if this association with CpG sites cg20321768 and cg25587058 is replicated in E-Risk participants using DNA methylation measured at age 18 and a similar facial emotion recognition task administered at age 10.

Statistical analyses:

Prior to analyses, DNAm and the facial emotion variable will be standardized. Associations will be tested using linear regression with correct detection of anger in 100% morphed angry faces from the Faces Game as the dependent variable (reversed prior to analysis so the direction corresponds to the error rate in ALSPAC) and DNA methylation at cg20321768 and cg25587058 as predictors (models run separately for each CpG site). Models will include covariates to adjust for cell count variation, sex, age at time of blood draw, and technical variation in DNA methylation measurements (e.g. plate/batch as necessary). Sensitivity analyses will evaluate associations in males only and further sensitivity analyses will include genetic variation (cis SNP rs138007199) as a covariate. All analyses will account for the non-independence of twin observations. For comparison with ALSPAC models, summary statistics will also be required for all covariates as well as the correlation between the two CpG sites.

Variables Needed at Which Ages (names and labels): Tony – please provide to Karen for analysis

Study: E-Risk

Age 5:

FAMILYID	Unique family identifier
ATWINID	Twin A ID (ex chkdig)
BTWINID	Twin B ID (ex chkdig)
RORDERP5	Random Twin Order
ZYGOSITY	Zygosity
SAMPSEX	Sex of Twins: In sample

Age 10:
CMANG100E10 - correct detection of anger at 100% morph

Age 18:

- DNA methylation levels at cg20321768 and cg25587058 from blood with 450k chip
- Cell count variation for samples used to generate DNA methylation profiles
- Genotypes for SNP rs138007199
- Technical variation in DNA methylation measurements (e.g. plate/batch)
- TAGEE18 Age at Interview - P18 – Elder

Data Security Agreement

Provisional Paper Title	FRAXA/E repeat length, DNA methylation and neurophenotypes in non-clinical populations
Proposing Author	Matthew Suderman
Today's Date	June 13, 2022

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

__ms__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/).

__ms__My project has ethical approval from my institution.

__ms__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.

__ms__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.

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

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

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

__ms__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.

__ms__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.

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

__ms__I will delete the data after the project is complete.

__N/A__ **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 <u>never</u> combined or stored with any other E-Risk data (family or twin-level data)

__ms__**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: Matt Suderman