**Chapter 3 Appendix A:**

[OSCA v0.45](https://yanglab.westlake.edu.cn/software/osca/#Overview) is a software used to analyse available omics (e.g. DNA methylation) data (Zhang et al, 2019). Three separate equations are provided to compute Omics Relationship Matrices (ORMs). In this study, we refer to ORMs as environmental relationship matrices (ERMs), as this is more appropriate for the data being used in this study (trauma principal components).

Algorithm 1 computes similarity by using standardised DNAm measures of all probes. Here it is important to note that the DNAm measures will be replaced with the principal components (PCs) of our trauma exposure measures.

With being the environmental similarity between individuals and . Where and is the **standardised** in individual and , respectively. and are the mean and variance of over all the individuals respectively.

Algorithm 2 computes similarity by using **unstandardised** PCs and factors in the number of PCs utilised. This algorithm implicitly assumes PCs with smaller variance tend to have larger effects on the phenotype, and that there is no relationship between the depression/neuroticism phenotypic variance captured by the PC and the variance of the PC. Here, we cannot confirm these assumptions, so equation 2, is deemed inappropriate for the data we have.

With being the environmental similarity between individuals and . Where and is the **unstandardised** in individual and , respectively. and are the mean and variance of over all the individuals respectively and is the number of PCs utilised.

Algorithm 3 computes similarity by iteratively standardising PCs and across individuals.

With being the environmental similarity between individuals and . Where and is the **standardised** in individual and , respectively. and are the mean and variance of over all the individuals respectively.

For our study, all algorithms are separately used to compute ERMs of full trauma PCs, algorithm 1 and 3 are deemed the most appropriate for our data. Subtle differences are observed in outputted relationship matrices (**Figure 1**). Matrices computed using algorithm 1 present diagonal values with a mean of 1, whereas matrices computed using algorithm 3 present diagonal values of 1.

|  |  |
| --- | --- |
|  |  |
| ***Figure 1.*** Trauma Exposure Relationship Matrices. Matrix diagonals represent similarity between an individual with themselves. Offdiagonals represent pairwise similarity between individuals within the sample. Here, similarity is computed using available trauma exposure principal components using OSCA software. Left, shows a representative similarity matrix where similarity is computed using Algorithm 1. Right, shows a representative similarity matrix where similarity is computed using Algorithm 3. Please note values specified are not obtained from real matrices. | |

Offdiagonal values of the computed ERMs and GRMs are plotted to visualise the relationship between values (**Figure 2**).

|  |  |
| --- | --- |
|  |  |
| ***Figure 2*.** Offdiagonal Values of UKB North Geographical Cluster Genomic and Environmental Relationship Matrices. Left, Environmental Relationship Matrix (ERM) computed using Algorithm 1 of OSCA software. Right, ERM computed using Algorithm 3 of OSCA software. The x-axis represent offdiagonal genomic similarity values, cluster points surrounding points 0.2-0.3 and 0.5 show points of relatives. The y-axis represents offdiagonal trauma exposure similarity values. | |

Zhang and colleagues (2019) demonstrated, with simulation analyses, that mixed linear model anlaysis results obtained from matrices computed with algorithm 1 and 3 showed minimal differences. Similarly, we show that GREML outputs were negligibly different between the algorithms of interest (appendix table 7).

Here, algorithm 1 is opted when computing the ERMs used for downstream analyses. The mixed linear model results show negligible differences and the plots from **figure** suggest values in line with expectations.

**Chapter 3 Appendix B:**

A consistent pattern observed in results are large LRT values when including ERMs into the mixed linear models (see **Table 1**). This suggests a substantial improvement in model fit when including the ERMs. Whilst the estimates and standard errors observed for the proportion of variance explained by the ERMs suggest statistical significance, they do not suggest very *strong* significance, which is discrepant with what the LRT values suggest.

|  |
| --- |
| **Table 1.** North Cluster Mixed Linear Model Results |
| |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | |  |  |  | **NORTH CLUSTER** | |  |  | | **Trait** | **Model** | **Source** | **Variance** | **SE** | **LRT** | **LRT-P** | | **CIDI** | **G** | **G** | 0.229 | 0.044 | 28.956 | 0.000 | | **E** | **E** | 0.170 | 0.055 | 1801.636 | 0.000 | | **G + E** | **G** | 0.171 | 0.040 |  |  | |  | **E** | 0.169 | 0.055 | 1792.306 | 0.000 | | **G x E** | **G** | 0.174 | 0.039 |  |  | |  | **E** | 0.176 | 0.057 |  |  | |  | **GxE** | 0.241 | 0.021 | 159.722 | 0.000 | |
| *Abbreviations.* CIDI, Composite International Diagnostic Interview Depression; G, Genomic Relationship Matrix; E, Environmental Relationship Matrix; GxE, Genome-by-Trauma Exposure Relationship Matrix; LRT, Log-Likelihood Ratio Test; P, P-value. |

To explore this discrepancy the log-likelihood distribution is examined. A log-likelihood function aims to identify and fit the most appropriate distribution to the data available (e.g. a normal distribution). A value from the observed data will be tested as the mean of the explored distribution. A log-likelihood value, signifying the likelihood of observing the data is obtained. A range of values are tested, and the resulting log-likelihood values are plotted to form the log-likelihood distribution. The highest value i.e. the maximum likelihood estimate, signifies the optimal position of the distribution explored to the data at hand.

Here, the discrepancy suggests a potentially unusual log-likelihood distribution which may be due to the nature of the ERMs used. Large LRT values would suggest the distribution to have a steep increase, whereas, the smaller standard errors would suggest a plateau surrounding the maximum estimate of the log-likelihood. To plot the log-likelihood distribution, a range of ERM values and their corresponding log-likelihood value was plotted (**Figure 3**).

|  |
| --- |
|  |
| ***Figure 3*.** Log-Likelihood Distribution of Mixed Linear Model Including the Trauma Exposure Environmental Relationship Matrix (ERM) for the UK Biobank North Geographical Cluster and Neuroticism Phenotype. The x-axis represents values of the ERM. The y-axis represents log-likelihood values. The orange point is the maximum likelihood estimate. |

The distribution observed explains the discrepancy observed between the LRT values and the estimates/standard errors of the variance attributable to the ERM.

**References:**

Zhang, F., Chen, W., Zhu, Z., Zhang, Q., Nabais, M. F., Qi, T., Deary, I. J., Wray, N. R., Visscher, P. M., McRae, A. F. & Yang, J. (2019) OSCA: a tool for omic-data-based complex trait analysis. *Genome Biol*, 20(1), 107.