Genome Wide Association Meta-Analysis of Math disabilities

Quality control of trait and genetic data & Analysis Plan (27 July 2021)

1. Background

This study aims to coordinate the collection of data for genome wide association meta-analysis (GWAMA) of quantitative mathematical/arithmetical traits. We have tried to keep the analyses to a minimum and tried to provide code and detailed instructions. We will provide assistance if requested. This protocol follows what used by Else Eising's for the GWAMA of Quantitative Reading and Language traits.

1.1 Aims

This analysis plan aims to i) coordinate the collection of phenotype descriptive statistics for the standardizing of the phenotypes, ii) coordinate the genotype quality controls and iii) coordinate the analyses.

1.2 Contact details

Questions about this analysis plan and resulting phenotype descriptions can be emailed to Filippo Abbondanza, fa36@st-andrews.ac.uk

2. Phenotype details

2.1 General comments

Make sure that your measure is positively measured, i.e. higher numbers reflect better performance.

2.2 Descriptions of phenotypes of interest

Looking at the traits available for each cohort, we propose the following phenotypes to be used for the GWAMA

Cohort Name	Phenotype	Predicted sample size	Tool for GWAS
Hong-Kong	Arithmetic	800	GEMMA
Generation R	CITO Math score	1800	GEMMA
NTR	CITO Math score	2000	GEMMA
FIOLA	Combined score of • Arithmetic fluency plus	400	GEMMA

	• Arithmetic fluency minus		
TEDS	Combined score of: • Understanding numbers • Computation and knowledge	4,000	BOLT-LMM
Toronto	WRAT	600	GEMMA
IOWA	Math computation	300	GEMMA
NeuroDys	HRT arithmetic subtests	800	GEMMA
Raine	MA	700	GEMMA
York	Combined score of: • OMA • OMS	200	GEMMA
ALSPAC	Paper 1 k3 test	10000	BOLT-LMM
Marburg Würzburg	Combine score of: • Addition verification • Multiplication verification	400	
Predicted sample size: 22,000			

The combined scores are the means between the standardised phenotypes highlighted in the table above

2.3 Phenotype description

Provide the following descriptive for each phenotype, performance IQ and age:

- histogram
- number of samples
- mean and median
- standard deviation and standard error of the mean
- skewness/kurtosis
- minimum and maximum values
- quartiles
- correlations between phenotypes
- correlation with age
- QQ plot

Provide the descriptive for all different measures (if available):

- for the raw measure
- for the scaled age-normalized measure

If phenotypes have been collected at multiple time points per individual, provide the descriptive of these phenotypes separately for each collection time point.

The provided R script `Pheno_analysis.R` can be used to obtain the descriptive. Please report back all tables and plots produced to fa36@st-andrews.ac.uk. This script provides the descriptions for males and females separately, as well as for the combined cohort.

Most cohorts have already provided this information, in which case no further action is required.

Note: All collaborators who will create composite scores as phenotype for the analysis, please run the `Pheno_analysis.R` to provide updated distributions.

3. Genotype data handling

3.1 Instructions for pre-imputation QC

We assume genotyping data has already gone through extensive quality control. If you have followed Else Eising's GWAMA for Quantitative language traits you should have already performed the following steps based on standard approaches, in which case no further action is required.

If you need to run QC, please exclude SNPs based on:

- Minor allele frequency (< 1%)
- Call rate (< 95% or 98%)
- Hardy-Weinberg (HWE p-value < 1×10^{-6})

Typically, studies have removed subjects based on:

- Sex inconsistencies based on X chromosome genotype data
- Call rate (< 95% or 98%)
- Heterozygosity
- Relatedness (IBD > 0.125, for samples with unrelated individuals only)
- Non-European ancestry based on PCA based analysis of genetic diversity

Only samples with European ethnicity, as identified using PCA-based analysis of genetic diversity, should be included in the imputation step. We will contact admixed non-European cohorts separately to discuss their sample filtering and imputation strategy.

3.2 Instructions for imputation

Genotypes should be imputed against reference panel HRC r1.1 (phased, hg19/build37). Imputation can be performed via the Michigan imputation server (https://imputationserver.sph.umich.edu/), the Sanger imputation server (https://imputation.sanger.ac.uk/) or manually. Please use the phasing of the HRC reference.

Post-imputation QC is not required; this will be performed at the meta-analysis stage.

4. Analysis plan

4.1. Phenotype data transformation and outliers

All phenotypes will need to be age-adjusted and z-standardised. For each phenotype outliers should be removed. This will be at the discretion to each cohort, but normally outliers were removed for scores >+-4SD.

4.2. Association analysis strategy

Analyses will be carried out in each cohort for:

- Combined sex
- Males-only (if enough sample size is available, ie. N > 150)
- Females-only (if enough sample size is available, ie. N > 150)

Phenotype data should be adjusted for sex when analysing both sexes together. As linear mixed models will be used for the analyses, there is no need to include principal components (PCs) as covariates.

The analysis will only include autosomal chromosome at this stage. X chromosome will be covered in a follow-up study

An example script for imputed data in BIMBAM format (preferred) is provided below for the cohorts using GEMMA software.

Minimac3/4 vcf files from the Michigan imputation server can be converted to plink dosage format using DosageConvertor (https://genome.sph.umich.edu/wiki/DosageConvertor), and subsequently to BIMBAM format by changing the field separators into commas. An example script is given below.

PBWT vcf files from the Sanger imputation server can be converted to gen files using bcftools (https://samtools.github.io/bcftools/). Gen files can easily be transformed into BIMBAM format using the awk script below.
Several other file formats can be converted to BIMBAM format using fcGENE software (http://www.bx.psu.edu/~giardine/tests/tmp/fcgene-1.0.7.pdf).

Scripts to run GWAS using BOLT/GEMMA are available at
https://github.com/fabbondanza/GenLang_Math_GWAS
Note: those scripts will have to be customised for your needs, they are intended to guide you in the process in case you are not familiar with GEMMA and/or BOLT.

4.3 Example scripts

Prepare BIMBAM files from Minimac3/4 vcf files using DosageConvertor

```
./DosageConvertor \
--vcfDose Imputed.chr1.dose.vcf.gz \
--info Imputed.chr1.info \
```

```
--prefix Imputed.chr1 \
--tag GP \
--type plink \
--format 1
gunzip Imputed.chr1.plink.dosage.gz
tr '\t' ',' Imputed.chr1.plink.dosage > Imputed.chr1.bimbam
awk 'BEGIN {OFS = ","} {print $2, $4, $1}' Imputed.chr1.plink.map > Imputed.chr1.map
```

Prepare BIMBAM files from gen files using awk

```
cat Imputed.chr1.gen | awk -v s=[number of individuals] '{ printf $2 "," $4 "," $5; for(i=1;
i<=s; i++) printf "," $(i*3+3)*2+$(i*3+4); printf "\n" }' > Imputed.chr1.bimbam
```

4.4 Example scripts for analysis in GEMMA

Estimate relatedness matrix with GEMMA **using pruned SNPs** (eg. Which can be generated with PLINK using --indep-pairwise 200 10 0.1 and –maf 0.1)

```
./gemma.linux \
-g Pruned.all_chr.bimbam \
-p phenotype_file.txt \
-gk 1 \
-o relatedness_matrix
```

Association analysis in samples with related individuals with Gemma

```
./gemma.linux \
-g Imputed.chr1.bimbam \
-a Imputed.chr1.map \
-p phenotype_file.txt \
-k relatedness_matrix.cXX.txt \
-lmm 1 \
-o WR_chr1
```

5. Results

5. Results

Please submit a tab-delimited summary file for each association analysis.

5.1 Requirements

1. SNP positions should be based on the NCBI build 37 map.
2. Report results for all autosomes in a combined file.
3. Data should be aligned to the plus (forward) strand.
4. Missing values should be coded as 'NA' (without quotes).
5. No quotes should be used around any data cells or headers

6. Please provide results for all imputed SNPs regardless of imputation quality or minor allele frequency.
7. Please report exact, unadjusted p-values; do not include 0 as p-value.
8. Please report at least 4 digits after the decimal place for all statistics.
9. No row indices but column headings should be provided.

5.2. Result format

Please use the following format for your result files

Column header	Description	Example
SNPID	SNP rs number	rs123456
CHR	Chromosome	1
POS	bp position	895234
STRAND	The strand on which the alleles are reported: we request to the forward (+) strand.	+
EFFECT_ALLELE	Allele for which the effect (BETA) is reported	T
NON_EFFECT_ALLELE	Other allele at this site	C
HWE_P	Exact P value of HWE test	0.4589
EAF	Effect allele frequency	0.1248
N	Number of samples analysed at this site	3860
BETA	Effect size	0.00369
SE	Standard error of BETA	0.00017
PVAL	Uncorrected two-sided P value	0.1269
IMPUTED	'0' for genotyped and '1' for imputed SNPs	1
INFO_TYPE	Imputation accuracy measure: 0 = genotyped 1 = BEAGLE (allelic r2) 2 = IMPUTE2 (info) 3 = MaCH (r2) 4 = Minimac (Rsq) 5 = Other	2
INFO	Selected imputation accuracy measure	0.8963

5.2. File naming

Please provide all files from your study named according to the following naming scheme:

COHORT_SEX_DATE.

COHORT refers to abbreviated cohort name
SEX refers to "MALE", "FEMALE" or "COMBINED"
DATE is the date on which the file was prepared, in the format YYYYMMDD.

Additionally, please provide details on the pre-imputation QC and imputation in the spreadsheet, if not already done so, to the following file:

https://docs.google.com/spreadsheets/d/1mikWXKxqgh9P4QJpOfMoDlIcu3OdA-6A/edit?usp=sharing&ouid=104947410872810207731&rtpof=true&sd=true

5.4. Results upload

Once the results are ready, please email fa36@st-andrews.ac.uk to organise the data transfer