ACTION GWAS Aggressive behavior and Attention Problems

This document details a standard operating procedure (SOP) for the GWA analyses that will be performed on Aggressive behavior (AGG) and Attention problems (AP) in each of the participating cohorts.

Contact details
Questions about this SOP can be directed to:
Michel Nivard, m.g.nivard@vu.nl
Meike Bartels, m.bartels@vu.nl
Dorret Boomsma, di.boomsma@vu.nl

Please CC all three of the contact persons in your email.

1 Background and Summary

The goal of this initiative is to carry out a meta-analysis of GWAS on Aggressive behavior (AGG) and Attention problems (AP) in subjects between the ages of 3 and 18. To attain this goal, phenotype data collected at different ages and rated by different raters (paternal, maternal, self and / or teacher) will be included in a single GWAS meta-analysis. The inclusion of multiple measures obtained on the same individuals, possibly including subjects rated by more than one rater and more than one age is somewhat more involved than for a single phenotype. However, for the participating cohorts the analyses consist of running a (possibly large) series of univariate analyses, while the inclusion of repeated measures and measures based on multiple informants will be done at the MA level, thereby substantially increasing power, while still allowing for the expression of age or rater dependent genetic effects.

The approach for the cohorts is as follows:
Every cohort will run a separate (univariate) analysis for each phenotype (i.e. one analysis for every combination of rater, age bin and survey measure). On the meta-analysis level, the test statistic per SNP will be combined over rater and age bin, accounting for the dependence between measurements (e.g. the same children rated by father and mother or at multiple ages).

To facilitate this integration, we request that each cohort supplies

- The covariance between the different measures (e.g. between raters, or between ages)
- A quantification of sample size of each age and rater bin
- The sample overlap between age and rater bins.

Our design allows us to identify to what extent the SNP effect changes with age, instrument or with the rater of the behavior. If, on the other hand, no robust associations are uncovered, it will allow us to put a much tighter upper bound on the expected effect sizes for common variants, and tight bounds on the variance in effect size found over age and or rater. Therefore, the results will have scientific merit regardless of the outcome.

For each cohort, a cohort-specific SOP will be developed based on information provided by the cohort on their specific phenotypes and in close consultation with the responsible PI / analyst.
For ACTION cohorts the SOPs will be based on the information ACTION has on the available AGG & AP related phenotypes in each cohort and consultation with PIs.
For the EAGLE cohorts we will base the sub-SOP on the recent EAGLE grant application efforts and consultation with PIs.
Additional information required from cohorts will include statistics based on genotyped individuals who will be included in the GWAS and meta-analysis.

2 Phenotype Definition

Our aim is to be inclusive with respect to the age range of subjects included in the analyses. We will include AGG and AP measures obtained for subjects between the ages of 3 and 18. We will further accept ratings based on a variety of psychometric instruments (e.g. CBCL, ASR, YSR, SDQ) and further encourage cohorts to contribute measurements obtained from different informants (self, mother, father and/or teacher). Our analysis strategy will allow inclusion of repeated measures and multiple raters, and can account for age-dependent genetic effects.

We take AP, or attention problems to be defined as a broad measure of attention and hyperactivity symptoms. So please consider using a broad scale which measures both inattention and hyperactivity if you have separate subscales available for attention and hyperactivity.

The sample-size threshold for inclusion is 1000 subjects for at least 1 rater at a single age. If this threshold is met, assessments of AGG and AP by any other rater or at another age in the same cohort can be included in the meta-analysis if the sample size at this measurement exceeds 500 subjects. These cutoffs are necessary as we assume the test statistics obtained for each SNP to be approximately normally distributed, but this is only true if a sufficiently large sample is analyzed in the individual cohorts.

3 Instructions for AGG and AP phenotypes

Transformation of the phenotype
We ask that the AGG or AP phenotypes are **not** transformed before analysis. To allow of the Beta's prior to meta-analysis we ask the analysis reports summary statistics per trait as outlined above.

The decision not to transform is based on a power analysis which compares the relative performance (in terms of power and type-1 error rate) of analyzing moderately to severely skewed data using either a quassi-poisson, logistic (mean split to arrive at categorical phenotype), linear regression and transforming the skewed data using a square root transformation followed by linear regression. Logistic regression performed worse in terms of power while the other strategies performed about equally well. All strategies retained acceptable type-1 error rates.

Please make sure that AGG / AP is positively measured, i.e. **higher numbers = higher score**. Please reverse your measure if you have a scale where higher number = lower score.

Sample Inclusion
We propose to limit the analyses to subjects with European ancestry.

Report the following phenotypic measures on the relationship between phenotypes:
1. Compute and report the phenotypic covariance (not correlation) matrix between all the measures (i.e. between the phenotypes used in all GWAs for AP **and** or AGG)
2. Report the sample size for each cell in your covariance matrix (diagonal and off-diagonal)

Further metric to be reported for each phenotype:
1. Mean age of the subjects
2. Informant (Maternal, paternal, self, teacher or other (in case of other please specify the informant))
3. Percentage male participants
4. Scale name, items included in scale (in English)
5. Scale mean and variance (at least 5 decimal places)
6. Scale Skewness and Kurtosis

4 Instructions for genotype handling

Pre imputation QC
We assume genotyping data has already gone through extensive quality control. Typically, studies have excluded SNPs from further analysis (or imputation) with:

- Minor allele frequency <1%
- Call rate <95% (or <99% if SNP has MAF < 5%)
- Failure of HWE exact test at p< 1e-6
- Known to have evidence of poor clustering on visual inspection of intensity plots

Typically, studies have removed subjects that have:

- Low overall call rates (< 95%)
- Excess autosomal heterozygosity
- Duplicate samples
- Known 1st or 2nd degree relatives in the sample (i.e. leave only one from each pedigree) **(unless the association analysis is family-based and correctly takes into account the observed relatedness)**
- Wrong gender (excessive X-chromosome homozygosity in males)
- XXY genotype etc.

Imputation
Genotypes should be imputed to 1000 genomes Phase III release and coordinates as used in the human reference genome GRCh37.

Imputation can be performed on the Michigan Imputation server (https://imputationserver.sph.umich.edu/index.html) or the service provided by Sanger in Oxford (https://imputation.sanger.ac.uk/).

Alternatively, we can recommend either IMPUTE (www.stats.ox.ac.uk/~marchini/software/gwas/gwas.html) or MACH (www.sph.umich.edu/csg/abecasis/MACH) imputation software. Please note that this is not a statement about the merits of these programs over other programs.

Please contact us if you have imputed genotypes to a different reference set or imputation mapped to a differenced human genome reference (for example GRCh38).

5 Instructions for Association tests and genome-wide analysis

We require you to run the following linear regression model:

$$Phenotype = \beta_0 + \beta_1 * SNP + \beta_2 * Z(age) + \beta_3 * sex + \beta_{4-9} * PC's + e$$

Covariates coding

- sex (coded 0 = F, 1 = M)
- z-score of age at the time of assessment
- The first 5 principal components. If, however, specific cohort-level analyses or prior knowledge of the sample structure suggests more principal components are needed to control for stratification, please include more PCs.
- If necessary, add study-specific covariates such as study site or batch effects (please discuss / specify)

Association tests

Association is tested on the 22 autosomes only. We exclude the sex-chromosomes and mitochondrial DNA. The association test preferably accounts for genotype imputation uncertainty (as for example is done with MACH2QTL and SNPTEST). Please provide the per-SNP quality indicators of imputation (proper_info in IMPUTE, r^2.hat in MACH).

If your sample includes closely related subjects (i.e. twins, siblings, families):

We request using the appropriate software (e.g. GCTA or BOLT-LMM) to conduct GWAs with related individuals. Please use a mixed effects procedure that includes 2 GRM's (see Tucker et al. 2015).

Briefly this procedure is performed as follows:

1. Compute a Genetic relatedness matrix (GRM) based on genotyped SNPs (if this is difficult due to multiple genotype platforms being used in your cohort see: Fedko et al. (2015) for an alternative). In computing this GRM omit SNPs that fail HWE ($p<$ 1e-6), have low MAF (<1%) or a low call rate (<99%).

2. Change the GRM into a GRM containing the pairwise genetic relationships above 0.05 (changing the values below 0.05 to 0). This GRM will form the first GRM in the mixed model.
3. Compute a series of GRMs each of which omits a single chromosome. These will form the remaining GRMs for the mixed model.
4. Run a mixed model in which the SNP is associated with the phenotype, while the model accounts for 1) the covariates defined above, and 2) the random effects associated with the 2 GRMs. When running this model for SNPs on a given chromosome, please make sure to include 1) the GRM based on all SNPs (with values < 0.05 set to 0) and 2) the GRM that omits this specific chromosome.

Mixed model analyses as outlined above can be performed in several software packages:

GCTA (http://cnsgenomics.com/software/gcta/mlmassoc.html),
FastLMM (http://research.microsoft.com/en-us/um/redmond/projects/mscompbio/fastlmm/)
BOLT-LMM (https://data.broadinstitute.org/alkesgroup/BOLT-LMM/).

Given several cohorts provided feedback on this we wish to point you to the software tool "RAEMETALWORKER". This GWAS tool allows for a correction for related subjects, the direct use of VCF files, and it can analyze DOSAGE data, which is considered superior to using best-guess data.

NB: Genomic control will be applied centrally at the meta-analysis stage (so do not perform any corrections beforehand).

If you require any assistance or have questions on the specific procedures, please contact Michel Nivard (m.g.nivard@vu.nl).

6 Instructions for sample and genotype description

Sample description
State the total sample size of the originally genotyped sample, and then indicate how many subjects were excluded and why (failed genotype QC, non-European, no phenotype data).

For genotyping, indicate the platform used (e.g. Affymetrix 5.0, Illumina 1M, Perlegen 600K), the number of SNPs that survive pre imputation QC and a general description of the QC methods. Also indicate how imputation was done, including the reference build, and some metrics on the general success of imputation.

Please do not forget to report the phenotypic descriptions requested in section 3.

7 Instructions for reporting results from first-pass association analyses

On the joint data from all participating cohorts a meta-analysis will be performed on the study-specific association statistics. This requires each participating study to report the following characteristics for every SNP (*all* imputed and observed SNPs are to be reported,

i.e. no p-value cut-off, no imputation quality cut-off and no MAF cut-off) in plain-text ASCII files for each phenotype separately:

Variable name (case sensitive!!)	**Description**
SNPID	SNP ID as rs number
Chr	Chromosome number (1-22).
position	physical position for the reference sequence (indicate build in readme file**, especially if this deviates from build 37**)
coded_all	Coded allele, also called modelled allele (in example of A/G SNP in which AA=0, AG=1 and GG=2, the coded allele is G)
noncoded_all	The other allele
strand_genome	+ or -, representing either the positive/forward strand or the negative/reverse strand of the human genome reference sequence; to clarify which strand the coded_all and noncoded_all are on
Beta	Beta estimate from genotype-phenotype association, at least 5 decimal places – 'NA' if not available
SE	Standard error of beta estimate, at least 5 decimal places – 'NA' if not available
Pval	*p*-value of test statistic, here just as a double check – 'NA' if not available
AF_coded_all	Allele frequency for the coded allele – 'NA' if not available
HWE_pval	Exact test Hardy-Weinberg equilibrium *p*-value – only directly typed SNPs, NA for imputed
callrate	Genotyping call rate after exclusions
n_total	Total sample with phenotype and genotype for SNP
imputed	1/0 coding; 1=imputed SNP, 0=directly typed
used_for_imp	1/0 coding; 1=used for imputation, 0=not used for imputation
oevar_imp*	Observed divided by expected variance for imputed allele dosage – NA otherwise
avpostprob**	Average posterior probability for imputed SNP allele dosage (applies to best-guess genotype imputation)

* oevar_imp is called r^2 in Mach, proper_info in Impute and R^2 in Beagle.
** avpostprob is called Quality in Mach, certainty in Impute and Beagle does not give this statistic.

Please upload a README file with a very brief description of the data uploaded, the date, the NCBI human genome reference sequence used for strand reference, and the scale of the beta estimates.

All numeric data can be specified in decimal notation, with at least 6 digits precision after the decimal dot. Please make sure that scientific notation is used for low p-values, e.g. the format 1.02E-07. Code missing values in any column as NA. No quotes should be used around any data cells or headers. Please provide columns labels in the first row of the file.

8 Instructions for uploading data

The results from the association analyses can be uploaded using secure file transfer protocol (sftp).

Host: lisa.surfsara.nl

User: action

Password: FP7_action

We will set up a common folder for all uploads instead of having to generate a folder for each IP address. The content of the common folder will be automatically migrated to another secure location every 24 hours.

To upload the data you need a sftp program. Both filezilla and winscp are freely downloadable (filezilla-project.org/ ; winscp.net/eng/download.php).

The following naming scheme for the files with your association results is preferred:
STUDY.PHEN.DATE.txt

where,

STUDY is a short identifier for the cohort studied
PHEN refers to phenotype
DATE is the date on which the file was prepared, in the format DDMMYYYY

9 Meta-analysis

Meta-analysis of results of all samples will be carried out by Hill Fung Ip and Koen Bolhuis and be supervised by Michel Nivard. We will apply genomic control based on the LDscore intercept (Bullik-Sullivan et al. 2014) and the appropriate marker filters at this stage (therefore, please provide unfiltered results).

As the meta-analysis will take the form of a multivariate meta regression (see http://biorxiv.org/content/early/2016/05/11/052829) it is of the utmost importance to accurately report the covariance between the individual AGG and AP phenotypes *within* cohort as well as the sample size and sample size overlap between the within cohort AGG and AP phenotypes. Integrating repeated measures and multiple raters will greatly benefit

power. Take, for example, the NTR dataset where up to ~2700 children are phenotyped or genotyped at 7 age points (3, 7, 10, 12, 14, 16 and 18 years) by either their mother, father or a teacher, this would yield 2700 independent observations to be analyzed in a univariate meta-analysis, while 22.000+ observations are available.

By performing univariate analyses and combining all results in meta-analysis, a loss of data is avoided, and an increase in power is realized. Moreover, this approach allows for a formal test of rater, instrument / diagnosis, age and cohort interaction effects.

These meta-analysis models allow us to not only account for age, rater or instrument effects while increasing power, but also to interpret and disseminate these age and rater specific effects to the broader field of psychiatric genetics. It is of clinical and translational interest to know whether genome-wide hits for adult neurological or psychiatric traits have an effect in childhood. If such an effect is present it is of further interest whether it increases with age or is present from early childhood onward. The results of our developmentally sensitive GWAS could thus be used to enrich and enhance GWAS analysis of adult phenotypes. The EAGLE and ACTION cohorts are uniquely able to perform these age and rater sensitive analyses given the repeatedly measured data they have aggregated.

10 References

Bulik-Sullivan, B. K., Loh, P. R., Finucane, H. K., Ripke, S., Yang, J., Patterson, N., ... & Schizophrenia Working Group of the Psychiatric Genomics Consortium. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics*, *47*(3), 291-295.

Fedko, I. O., Hottenga, J. J., Medina-Gomez, C., Pappa, I., van Beijsterveldt, C. E., Ehli, E. A., ... & Middeldorp, C. M. (2015). Estimation of Genetic Relationships Between Individuals Across Cohorts and Platforms: Application to Childhood Height. *Behavior genetics*, *45*(5), 514-528

Minică C. C., Boomsma D. I., Vink J. M., Dolan C. V. M. Z. (2014) twin pairs or MZ singletons in population family-based GWAS? More power in pairs. *Mol Psychiatry,* 19(11), 1154-5.

Minică C. C., Dolan C. V., Kampert M. M. , Boomsma D. I., Vink J. M. (2015) Sandwich corrected standard errors in family-based genome-wide association studies. *Eur J Hum Genet.* 23, 388-94.

Pappa I, St Pourcain B, Benke K, Cavadino A, Hakulinen C, Nivard MG, Nolte IM, Tiesler CM, Bakermans-Kranenburg MJ, Davies GE, Evans DM, Geoffroy MC, Grallert H, Groen-Blokhuis MM, Hudziak JJ, Kemp JP, Keltikangas-Järvinen L, McMahon G, Mileva-Seitz VR, Motazedi E, Power C, Raitakari OT, Ring SM, Rivadeneira F, Rodriguez A, Scheet PA, Seppälä I, Snieder H, Standl M, Thiering E, Timpson NJ, Veenstra R, Velders FP, Whitehouse AJ, Smith GD, Heinrich J, Hypponen E, Lehtimäki T, Middeldorp CM, Oldehinkel AJ, Pennell CE, Boomsma DI, Tiemeier H. A genome-wide approach to children's aggressive behavior: The EAGLE consortium. *Am J Med Genet B Neuropsychiatr Genet.* 2016, 171(5):562-72

Tucker, G., Loh, P. R., MacLeod, I. M., Hayes, B. J., Goddard, M. E., Berger, B., & Price, A. L. (2015). Two-Variance-Component Model Improves Genetic Prediction in Family Datasets. *The American Journal of Human Genetics*, *97*(5), 677-690