4 December 2013

Gentrepreneur Consortium Analysis Plan

Philipp Koellinger, Niels Rietveld, and Rotterdam Study Entrepreneurship GWAS working group

1 Motivation

Economic variables such as income, education, and occupation are known to affect mortality and morbidity, and have also been shown to be partly heritable. However, very little is known about which genes influence economic variables, although these genes may have both a direct and an indirect effect on health. Therefore, Van der Loos et al. (2013) studied the molecular genetic architecture of an economic variable–entrepreneurship, which was operationalized using self-employment, a widely available proxy. Although the results suggest that self-employment is a polygenic trait, a meta-analysis of genome-wide association studies across seventeen studies comprising 53,898 participants did not identify genome-wide significant SNPs. The aim of this project is to increase the sample size of the meta-analysis to enable identification of SNPs associated with self-employment with very small effect sizes. The research objective is twofold: 1) identification of association SNPs and self-employment by applying meta-analysis of GWAS from different cohorts and studies using imputed SNP data, and 2) construction of a polygenic risk score using the meta-analysis results to predict self-employment out of sample.

2 Gentrepreneur Consortium

The Gentrepreneur Consortium is a combination of the cohort studies participating in the Working Group (WG) on entrepreneurship within the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium and additional cohort studies recruited to increase power (see Van der Loos et al., 2010, for an extended description of the consortium set-up). The aim of the consortium is to identify loci associated with self-employment by applying meta-analysis of GWAS of different cohort studies using imputed SNP data.

3 Trait definitions

Self-employment: At least once self-employed (0 = Never self-employed, 1 = At least once self-employed).

Serial self-employment: More than once self-employed; meaning the individual had more than one spell of self-employment in her or his work-life history (0 = Never self-employed, 1 = At least twice self-employed).
→ Define and analyze this trait only if your dataset permits.

Only self-employment: Only self-employed; meaning the individual had possibly more than one spell of self-employment in her or his work-life history, but was nevertheless always self-employed in her or his work-life (0 = Never self-employed, 1 = Only self-employed). → Define and analyze this trait only if your dataset permits.

NOTE: The definitions of self-employment are nested, meaning that the at least once self-employed group will *also* contain serial and only self-employed individuals. Additionally, the serial self-employed group will also contain only self-employed individuals. Only self-employed is the most narrow definition of self-employment and this group will *just* contain only self-employed individuals.

Control group: The control group in the analyses should ALWAYS be those individuals who were never self-employed in the past (and ideally will not be in the future), i.e. the control

group should not contain individuals who have been self-employed at some point. Incomplete work-life histories can lead to control groups that contain individuals who will or already have been self-employed, which will decrease the power of the statistical analysis and make the interpretation of results difficult.

NOTE: Please exploit all available information in your dataset to define the control group and the self-employment variables as precisely as possible and double-check your definition with Philipp Koellinger (koellinger@ese.eur.nl) before running GWAS.

4 Genotypes & imputation

All autosomal SNPs imputed from the HapMap Phase II CEU panel to allow analyses across different genotyping platforms (Affymetrix, Illumina, Perlegen).

Recommended marker filters to be applied before imputation: SNP call > 95%, HWE $p > 10^{-6}$, MAF > 5%.

The following marker exclusions will be applied *at* the meta-analysis stage: proper_info < 0.40 (SNPtest), r2.hat < 0.30 (MACH).

Please provide *unfiltered* results, as these filters will be applied at the meta-analysis stage.

5 Population stratification

To control for population stratification the first four principal components (PCs) should be included in the individual analyses as covariates. Inclusion of the PCs is conditional upon availability within individual studies.

Genomic control (GC) will be applied to each study *at* the meta-analysis stage (single GC). Second, overall GC will be applied to the meta-analysis results (double GC). Double GC adjusted results will be compared to single GC adjusted results, as important results may be discarded due to the over-conservative nature of double GC.

Please do *not* apply GC to GWAS results and provide uncorrected standard errors, as (double) GC will be applied at the meta-analysis stage.

6 Model used to test for association

The model to be used is an additive logit model with a dummy for sex and dummies for different age categories. Age is defined as the age at time of survey of the occupational status variable or the variable(s) used to code the self-employment phenotype. Age categories are defined as:

- <= 29 (reference);
- 30-39;
- 40-49;
- >= 50.

Depending upon availability, include the first four principal components as covariates in the logit model.

The full model is thus specified as:

$$\text{logit}(p) = \beta_0 + \beta_1\text{SNP} + \beta_2\text{Sex} + \beta_3\text{AgeCat2} + \beta_4\text{AgeCat3} + \beta_5\text{AgeCat4} \; (+ \beta_6\text{PC1} + \beta_7\text{PC2} + \beta_8\text{PC3} + \beta_9\text{PC4})$$

We encourage analysing genotypes as allele doses (to account for imputation uncertainty).

7 Timeline for delivery of results

GWAS analyses by: **To be decided.**

If you have problems meeting this deadline please contact: Olga Rostapshova (olga_rostapshova@hksphd.harvard.edu).

8 Results file formats

8.1 File formatting

SNP Table 1 for association results

Results should be formatted according to the CHARGE results sharing format (http://depts.washington.edu/chargeco/wiki/ResultsSharingFormat).

Variable name (case sensitive!!)	Description
SNPID	SNP ID as rs number
chr	Chromosome number. Use symbols X, XY, Y and mt for non-autosomal markers.
position	physical position for the reference sequence (indicate build 35/36 in readme file)
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, to 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=if 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)

Please note that a README should be uploaded with a very brief description of the data uploaded, the date, the NCBI human genome reference sequence used (e.g. NCBI 36.2) for strand reference, and the scale of the beta estimates.

SNP Table 2 for top SNPs

This table will be requested for top SNPs only!

Variable name (case sensitive!!)	Description
callrate	SNP call rate: NTOT/NGENO
HWE_pval	Hardy-Weinberg *p*-value
n0	Number individuals homozygous major allele
n1	Number heterozygous individuals
n2	Number individuals homozygous minor allele
fu0	Trait frequency in individuals homozygous for major allele
fu1	Trait frequency in heterozygous individuals
fu2	Trait frequency in individuals homozygous for minor allele
r2	Regression *R*-squared

8.2 File naming scheme

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

SNP table 1: **STUDY**.**PHENOTYPE**.association-results.**DATE**.txt
SNP table 2: **STUDY**.**PHENOTYPE**.topsnps.**DATE**.txt

where,

STUDY is a short (14 characters or less) identifier for the population studied.

PHENOTYPE is one of:
'once' indicating at least once self-employed.
'serial' indicating at least twice self-employed.
'only' indicating always self-employed.

DATE is the date on which the file was prepared, in the format 'YYYYMMDD'.

9 Data exchange procedure

Summary statistics can be entered into an Excel spreadsheet, which has been circulated and is available upon request.

GWA results will be uploaded to the ShareSpaces file sharing system (http://www.washington.edu/lst/web_tools/sharespaces). Users that already have a user name, please inform Niels Rietveld (nrietveld@ese.eur.nl) with your user name (or the user name of your data analyst) so access can be granted. Otherwise, new accounts can be registered at the same website. After registration please also inform Niels Rietveld with your user name.

10 Meta-analysis

Meta-analysis will be performed using the inverse-variance methodology as implemented in the software METAL[1]. As mentioned above, we will apply (double) genomic control and the appropriate marker filters at this stage.

The fixed threshold for genome-wide significance is set at $p < 5 \times 10^{-8}$. Any marker reaching this threshold and showing a consistent relationship across the majority of the studies will be eligible for publication. We will define suggestive SNPs as $p < 10^{-6}$, but do not reach genome-wide significance. For markers reaching this threshold, we will possibly seek replication in additional cohorts.

11 References

Van der Loos, M.J.H.M., Koellinger, P.D., Groenen, P.J.F., & Thurik, A.R. (2010). Genome-wide association studies and the genetics of entrepreneurship. *European Journal of Epidemiology*, *25*, 1-3.

Van der Loos, M.J.H.M., Rietveld, C.A., Eklund, N., Koellinger, P.D., Rivadeneira, F., Abecasis, G.R., ... Thurik, A.R. (2013). *The molecular genetic architecture of self-employment*. PLOS ONE, 8(4), e60542.

[1] http://www.sph.umich.edu/csg/abecasis/metal/