GWAS on educational attainment

Philipp Koellinger

Erasmus School of Economics
Erasmus University Rotterdam

2nd SSGAC Workshop • 29 October 2011
Progress

- Established connections with
  - Over 50 GWAS cohorts
  - Major social science data providers (HRS, PSID, WLS)
- Database of available social science phenotypes
- Infrastructure and experience to facilitate large scale GWAS efforts
- Qualified meta-analysts
- http://www.ssgac.org
Progress

- Educational attainment:
  - Widely measured
  - Measures can be harmonized (ISCED)
  - Relevant in medicine and social sciences
  - Moderately heritable
    - Taubman 1976; Miller, Mulvey and Martin 2001
  - However, it’s biologically distal
- Analysis plan distributed in Feb 2011
- Deadline for uploading results was Jul 2011
- 5 conference calls
- 42 cohorts uploaded (N ~ 105,000)
Meta-analysis

• Analysts:
  – Niels Rietveld (Economics, Erasmus U Rotterdam)
  – Nico Martin (Queensland Institute of Medical Research)
  – Jamie Derringer (Psychology, U Minnesota)

• Methodological advise:
  – Sarah Medland (Queensland Institute of Medical Research)

• Quality control:
  – $MAF > 1\%$
  – Imputation quality $R^2 > 40\%$ (MACH and Impute)
  – Lambda < 1.05
  – Cohort-specific QQ and Manhattan plots
Issues

• A lot of follow-up work
• ¼ of the uploaded results looked unreasonable or indicated problems
  – Duplicate SNPs with different $p$-values
  – Inflated QQ plots (often small cohorts & low MAF)
  – Extremely low $p$-values for some SNPs in only one or two studies
• Results are preliminary and based on 80% of all cohorts (N ~ 85,000)
• Two model specifications
  – OLS on educational attainment in US schooling years (ISCED)
  – Logit on college degree
QQ plots educational attainment

females (edu_years)  
N = 51,333

males (edu_years)  
N = 52,601

pooled (edu_years)  
N = 32,601

females (college)  
N = 52,485  
Yes = 20%

males (college)  
N = 34,631  
Yes = 27%

pooled (college)  
N = 52,485  
Yes = 23%
Manhattan plots females

females (college), single GC, N = 52,485:

Genome-wide significance > 7.3

females (edu_years), single GC, N = 51,333:

Philipp Koellinger - Erasmus University Rotterdam
Manhattan plots males

males (college), single GC, N = 34,631:

Genome-wide significance >7.3

males (edu_years), single GC, N = 32,601:
Manhattan plots pooled

pooled (college), single GC, N = 87,116:

Genome-wide significance > 7.3

pooled (edu_years), single GC, N = 83,934:
Next steps

• Continue follow-up and QC
• Invite cohorts for replication stage
  – First in-silico
  – Then maybe wet-lab, if something replicates
  – Goal: N > 30,000 replication samples
• Additional analyses
Lessons learnt

• Effect sizes of common SNPs are very small
  – Top hits odds ratios: 0.9; 1.1
• Large $N$ and phenotype harmonization are important
• Looking at different proxies of the same phenotype and the tails of the phenotype distribution is a good idea
• QC, follow-up and logistic management take a lot of time
80% power calculations for college, given phenotype distribution and $N$