Syllabus summer school “Genome-wide Data Analysis” (2020) at Tinbergen Institute

Organization

The summer school takes place 6-10 July 2020 and is taught by Prof. Philipp Koellinger and Dr. Ronald de Vlaming from the Department of Economics at Vrije Universiteit Amsterdam. Guest lectures will be given by Dr. Fleur Meddens M.Sc. (Erasmus School of Economics, Erasmus University Rotterdam), Dr. Aysu Okbay (Department of Economics at Vrije Universiteit Amsterdam), and Dr. Abdel Abdellaoui (AMC Medical Research BV, University of Amsterdam).

All sessions will be held at the premises of the Tinbergen Institute (TI) in Amsterdam, The Netherlands, close to train station Amsterdam Zuid. The location can be easily reached from Schiphol airport (~15 min) and Amsterdam Central Station (~30min), and is within walking distance from the Vrije Universiteit Amsterdam (http://www.tinbergen.nl/contact/).

Every day, lectures take place from 09:30 – 12:30 and computer tutorials from 13:30 – 15:30. In addition, participants will be given a reading list for self-study and preparation as well as computer assignments for (guided) practice during the computer tutorials.

All participants are asked to bring their personal laptop to all computer tutorials. Free Internet access is available via EDUROAM (guest accounts will be available).

Short subject description

The goal of the summer school is to introduce researchers from various fields to key concepts, state-of-the-art methods, and computer tools in statistical genetics that can be applied in the social and medical sciences. The course will be highly quantitative and interactive, covering topics such as the estimation and interpretation of heritability using molecular genetic data, genetic association studies, polygenic prediction, and identification strategies to isolate causal effects using genetic insights. It will emphasize methodological issues such as appropriate study design, data integrity, multiple testing, detecting and controlling for potential confounds, as well as factors influencing the accuracy of polygenic scores.

Target group

The summer school welcomes (research) master students, PhD students, post-docs, and professionals from various disciplines (e.g. behavioral genetics, economics, medicine, sociology, political sciences, psychology) who are interested in learning state-of-the art methods for genome-wide data analysis.

The summer school will attempt to “bridge” specific knowledge gaps that participants from different backgrounds may have. Specifically, social scientists will get a short, formal introduction to genetics, while students from medicine or genetics will benefit from the formal treatment of statistical methods and the discussion of how investigating the genetics of social scientific outcomes may lead to medically relevant insights.

A formal background in statistics or econometrics is required from students (at the level of a first year course in a graduate school PhD program in economics, psychology, or epidemiology), but no formal background in genetics will be assumed. The course is taught in English and will be accessible to students from both in- and outside the Netherlands.

Evaluation

Participants who joined at least 80% of all sessions will receive a certificate of participation stating that the summer school is equivalent to a work load of 3 ECTS. Note that it is the student's own responsibility to get these credits registered at their own university.

Computers and software

Please bring your own laptop with you for the computer tutorials. During the tutorials, we will work with freely available software. In particular, we will use programming languages R and Python (Version 2.7). These languages are an important part of the toolkit of a data scientist.

For a smooth installation of R and Python, we recommend you to download and install Anaconda (https://www.anaconda.com/distribution/). Anaconda is an installation manager of sorts, that will help to get Python and R, as well as graphical user interfaces for these languages, up and running on your machine. When downloading Anaconda, please make sure to download the distribution that packs the **Python 2.7 version** (and not the one that packs Python 3.7)!

Once Anaconda is installed, you can use the now available Anaconda Navigator to install RStudio, a graphical user interface for R, from the home panel of the navigator. Installing RStudio this way will ensure that R is also automatically installed.

We also recommend to install Notepad++ (https://notepad-plus-plus.org/), a highly versatile text editor, and 7-Zip (http://www.7-zip.org/download.html), a file archiver supporting many compression formats. All other software tools that we will be using run without installation.

Social activities

Lunches will be provided in the TI building from Monday-Friday. On Tuesday evening we will have the summer school dinner (Bierfabriek, Nes 67, Amsterdam). There will be farewell drinks on Friday afternoon in the TI building.

Course outline

Note: References in **bold** are required reading, all others are optional (but highly recommended). Clicking on the references will take you directly to the articles (whenever possible, we used a link to an unrestricted version).

1) **Introduction (July 6)**
   a) Lecture – Part 1 (Koellinger)
      - Terminology
      - Overview of possible applications
      - Genetic data
   b) Lecture – Part 2 (Meddens)
      - Mendel’s laws of heredity
      - Exceptions to Mendel’s laws
      - Genetically complex traits
      - Hardy-Weinberg equilibrium
   c) Lecture – Part 3 (Meddens)
      - Linkage disequilibrium
      - Genotyping vs. sequencing
      - Interpreting the results of genetic association studies
   d) Computer tutorial (de Vlaming)
      - *Getting familiar with genetic data and PLINK*

Literature

Benjamin, D.J., et al. (2012). The promises and pitfalls of genoeconomics. *Annual Review of Economics*, *4*(1), 627–662.

Falconer, D.S., Mackay, R.F.C. (1995). *Introduction to Quantitative Genetics*. 4th edition. Pearson. Chapters 1-2. (on SurfDrive)

Freese, J. (2018). The arrival of social science genomics. *Contemporary Sociology*, *47*(5), 524-536.

Goldberger, A. (1979). Heritability. *Economica*, *46*(148), 327-347.

Jencks, C. (1980). Heredity, environment, and public policy reconsidered. *American Sociological Review*, *45*(5), 723-736.

Harden, K.P. and Koellinger, P.D. (2020). Social Science Genetics: New Methods for Enduring Questions. *Nature Human Behaviour*, forthcoming. (on SurfDrive)

Plomin, R., DeFries, J., Knopik, V.S., Neiderhiser, J.M. (2013). *Behavioral Genetics*. 6th edition. New York: Worth Publishers. Chapters 2-4. (on SurfDrive)

2) Molecular genetic basics and heritability (July 7)

a) Lecture – Part 1 (Koellinger)
- Conceptual framework for studying genetic effects on human traits

b) Lecture – Part 2 (Koellinger)
- Statistical power

c) Lecture – Part 3 (Koellinger)
- Broad- vs. narrow-sense heritability
- Interpreting heritability estimates
- Estimating heritability using molecular genetic data with GREML

d) Computer tutorial (de Vlaming)
- *R: The classical twin study*
- *Getting familiar with GCTA*

17:00: Dinner (Bierfabriek, Nes 67, Amsterdam)

Literature

Boyle, E.A., et al. (2017). An expanded view of complex traits: From polygenic to omnigenic. *Cell*, *169*(15), 1177-1186.

Chabris, C., et al. (2015). The fourth law of behavior genetics. *Psychological Science*, *24*(4), 304-312.

Evans, L.M. et al. (2018). Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits. *Nature Genetics*, *50*, 737-745.

Ioannidis, J.P.A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124.

Tam, V., et al. (2019). Benefits and limitations of genome-wide association studies. *Nature Review Genetics*, *20*, 467-484.

Moonesinghe, R., et al. (2007). Most published research findings are false – but a little replication goes a long way. *PLoS Medicine*, *4*(2), e28.

Visscher, P.M., et. al. (2017). 10 years of GWAS discovery: Biology, Function, and Translation. *American Journal of Human Genetics*, *101*, 5-22.

Visscher, P.M., Hill, W.G., Wray, N.R. (2008). Heritability in the genomics era — concepts and misconceptions. *Nature Review Genetics*, *9*(4), 255-266.

Yang, J. et al. (2010). Common SNPs explain a large proportion of the heritability of human height. *Nature Genetics*, 42(7), 565–569.

Young, A.I., et al. (2019). Deconstructing the sources of genotype-phenotype associations in humans. *Science*, *365*, 1396-1400.

3) Genetic discovery and population stratification (July 8)

a) Lecture – Part 1 (Koellinger)
- Candidate gene studies
- Genome-wide association studies (GWAS)
- Imputation
- Meta-analysis

b) Lecture – Part 2 (Abdellaoui)
- Population stratification
- Cryptic relatedness
- Genomic control
- Principal components

c) Lecture – Part 3 (Okbay)
- Quality control of GWAS results

d) Computer tutorial (de Vlaming)
- *GWAS with PLINK*
- *Population stratification correction with PLINK, GCTA, and R*
- *Visualizing GWAS results and quality control in R*

Literature

Abdellaoui, A. et al. (2013). Population structure, migration, and diversifying selection in the Netherlands. *European Journal of Human Genetics*, *21*(11), 1277-1285.

Devlin, B., Roeder, K. (1999). Genomic control for association studies. *International Biometric Society*, *55*(4), 997-1004.

Lin, D.Y., Zeng, D. (2009) Meta-analysis of genome-wide association studies: no efficiency gain in using individual participant data. *Genetic Epidemiology*, *34*, 60–66.

Marchini, J., Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, *11*(7), 499-511.

Mills, M.C. & Rahal, C. (2019). A scientometric review of genome-wide association studies. *Communications Biology*, *2*, article number 9.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., Reich, D. (2006). Principal component analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, *38*(8), 904-909.

Willer, C.J., Li, Y., Abecasis, G.R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*, *26*(17), 2190-2191.

Winkler, T.W., et al. (2014) Quality control and conduct of genome-wide association meta-analyses. *Nature Protocols*, *9*(5), 1192-1212.

4) **Genetic discovery – continued (July 9)**
   a) Lecture – Part 1 (Koellinger)
      - LD-score regression
   b) Lecture – Part 2 (Koellinger)
      - The endo- and proxy-phenotype approach
      - Multivariate analysis of traits
   c) Lecture – Part 3 (Koellinger)
      - Example: Educational attainment
   d) Computer tutorial (de Vlaming)
      - *Meta-analysis with METAL*
      - *LD Score Regression*

Literature

Bulik-Sullivan, B., et al. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics*, *47*(3), 291-295.

Bulik-Sullivan, B., et al. (2015). An atlas of genetic correlations across human diseases and traits, *Nature Genetics*, *47*(11), 1236-1241.

Grotzinger, A., et al. (2019). Genomic SEM provides insights into the multivariate genetic architecture of complex traits, *Nature Human Behaviour*, 3, 513-525.

Lee, J.L. et al. (2018). Gene discovery and polygenic prediction from a 1.1-million-person GWAS of educational attainment, *Nature Genetics*, *50*, 1112-1121.

Okbay, A., et al. (2016). Genome-wide association study identifies 74 loci associated with educational attainment, *Nature*, *533*(7604), 539-542

Rietveld, C.A., et al. (2013). GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *Science*, *340*(6139), 1467–1471**.**

Rietveld, C.A., et al. (2014). Common genetic variants associated with cognitive performance identified using the proxy-phenotype method. *Proceedings of the National Academy of Science of the United States of America*, *111*(38), 13790-13794.

Rietveld, C.A., et al. (2014). Replicability and robustness of genome-wide-association studies for behavioral traits. *Psychological Science*, *25*(11), 1975-1986.

Turley, P., et al. (2018). Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nature Genetics*, 50, 229-237.

5) Polygenic scores and applications (July 10)

a) Lecture – Part 1 (Koellinger)
   - Constructing polygenic scores
   - Accuracy of polygenic scores

b) Lecture – Part 2 (Koellinger)
   - Imperfect genetic correlation across samples in GWAS meta-analysis
   - Polygenic scores as control variables

c) Lecture – Part 3 (Koellinger)
   - Genes as instrumental variables (a.k.a. Mendelian Randomization)

d) Computer tutorial (de Vlaming)
   - *Constructing and working with polygenic scores in PLINK and R*
   - *Mendelian randomization in R*

15:30: Farewell drinks at Tinbergen Institute.

Literature

Bowden, J., Davey Smith, G., Burgess, S. (2015). Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *International Journal of Epidemiology*, *44*(2), 512-525.

Daetwyler, H.D., Villanueva, B., Wooliams, J.A. (2008). Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLoS One*, *3*(10), e3395.

Davey Smith, G., Hemani, G. (2014). Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Human Molecular Genetics*, *23*(R1), R89-R98.

De Vlaming, R. et al. (2017). Meta-GWAS Accuracy and Power (MetaGAP) calculator shows that hiding heritability is partially due to imperfect genetic correlations across studies. *PLOS Genetics*, *13*(1), e1006495.

Di Prete, T. et al. (2018). Genetic Instrumental Variable (GIV) regression: Explaining socioeconomic and health outcomes in non-experimental data. *Proceedings of the National Academy of Sciences of the USA*, 115(22), E4970-E4979.

Ge, T. et al. (2019). Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nature Communications*, *10*, 1776.

Koellinger, P.D. and de Vlaming, R. (2019). Mendelian randomization: the challenge of unobserved environmental confounds. *International Journal of Epidemiology*, https://doi: 10.1093/ije/dyz138.

Martin, A.R. et al. (2019). Current clinical use of polygenic scores will risk exacerbating health disparities. *bioRxiv* preprint, https://doi.org/10.1101/441261.

O'Connor, L.J., Price, A. (2018). Distinguishing genetic correlation from causation across 52 diseases and complex traits. *Nature Genetics*, *50*, 1728-1734.

Rosenberg, N.A. et al. (2019). Interpreting polygenic scores, polygenic adaptation, and human phenotypic differences. *Evolution, Medicine, an Public Health*, *1*, 26-34.

Verbanck, M. et al. (2018). Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nature Genetics*, 50, 693-698.

Vilhjálmsson, B. et al. (2015). Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *American Journal of Human Genetics*, *97*(4), 576-592.

Warren, M. (2018). The approach to predictive medicine that is taking genomics research by storm. *Nature*, *562*, 181-183 (News Feature).

Zhu, Z. et al. (2018). Causal associations between risk factors and common diseases inferred from GWAS summary data. *Nature Communications*, 9(224).