Statistical challenges in large-scale genetic studies

A research afternoon by the Centre for Statistics

March 26, 2024

James Clerk Maxwell Building, Room 5205

Schedule

12:00 Lunch and networking

12:45 Welcome

12:50 Keynote speakers

Ken Rice: *Statistical problems with GWAS – and some solutions for them*

Ignacy Misztal: *GWAS in large animal studies – why so few QTLs identified?*

13:50 Short talks

Sara Wade: *Leveraging variational autoencoders for multiple data imputation*

Ivan Pocrnic: *Making big data small and small data big: the story of genomic dimensionality*

Sjoerd Beentjes: *Semi-parametric efficient estimation of small genetic effects in large-scale population cohorts with TarGene*

Lijuan Wang: *Prioritization of therapeutic targets for cardiovascular diseases using integrative multi-omics and machine learning analyses*

14:40 Coffee break

15:00 Short talks

Ioannis Papastathopoulos: *Unveiling patterns in rare events: Statistical techniques for extreme multivariate events*

Michelle Luciano: *Multivariate genome-wide association of quantitative reading achievement scores and dyslexia diagnosis*

Aris Sionakidis: *Challenges in integrating high dimensional biomedical data*

Ismail Ozkaraca: *Divide and Conquer Approach in Genome-wide Association Studies*

15:50 Closing remarks

Abstracts

Statistical problems with GWAS – and some solutions for them

Ken Rice

High-throughput genotyping in large human cohorts has revolutionized genetic discovery work. However, despite the apparently-simple goal of these analyses – looking for variants associated with disease outcomes – a number of statistical difficulties crop up. In this talk, based on experiences in the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) and Trans-Omics for Precision Medicine (TOPMed) consortia, I will summarize several solutions for them, most of which are available in the GENESIS package.

GWAS in large animal studies – why so few QTLs identified?

Ignacy Misztal

An accurate GWAS requires a correct model, all applicable information including genotypes, pedigrees and phenotypes, and an appropriate computing technique. For populations with many genotyped individuals computing may be very expensive. In studies using farm animals, the dimensionality of the genomic information is low, below 20k. We developed a single-step GWAS methodology where the small dimensionality is exploited for drastically reduced computations. For large data up to half a million genotypes and one million phenotypes, our method found very few significant signals. We hypothesize that for traits under long term selection, nearly all large variances are near fixation, and the applied methodology eliminates spurious signals, e.g., due to genotyping errors (e.g., via imputation) and population structure.

Leveraging variational autoencoders for multiple data imputation

Sara Wade

Missing data persists as a major barrier to data analysis across numerous applications. Recently, deep generative models have been used for imputation of missing data, motivated by their ability to learn complex and non-linear relationships. In this work, we investigate the ability of variational autoencoders (VAEs) to account for uncertainty in missing data through multiple imputation. We find that VAEs provide poor empirical coverage of missing data, with underestimation and overconfident imputations. To overcome this, we employ β-VAEs, which viewed from a generalized Bayes framework, provide robustness to model misspecification. Assigning a good value of β is critical for uncertainty calibration and we demonstrate how this can be achieved using cross-validation. We assess three alternative methods for sampling from the posterior distribution of missing values and apply the approach to transcriptomics datasets with various simulated missingness scenarios. Finally, we show that single imputation in transcriptomic data can cause false discoveries in downstream tasks and employing multiple imputation with β-VAEs can effectively mitigate these inaccuracies

---

Making big data small and small data big: the story of genomic dimensionality

Ivan Pocrnic

Over the past decade, animal evaluation databases have amassed abundant genomic data, boasting millions of genotyped individuals and markers ranging from thousands to millions. However, the sheer scale of these datasets renders traditional quantitative genetic models ineffective, overwhelmed by their cubic computational complexity. We present the prevalent statistical models utilised in such contexts to springboard potential collaborations between statisticians and animal geneticists. We then explore computational techniques to enhance efficiency, emphasising leveraging the limited dimensionality of genomic data. Additionally, we touch the flip side of the coin: the numerous small databases requiring specialised handling for meaningful inference.

Semi-parametric efficient estimation of small genetic effects in large-scale population cohorts with TarGene

Sjoerd Beentjes

I will discuss a unified statistical workflow for the semiparametric efficient and double robust estimation of causal n-point interactions amongst categorical variables in the presence of confounding and weak population dependence. N-point interactions, or Interaction ATEs (IATEs), are a generalisation of the usual average causal effect. To estimate IATEs, we introduce cross-validated and/or weighted versions of Targeted Minimum Loss-based Estimators (TMLE) and One-Step Estimators (OSE). The effect of dependence amongst units on variance estimates, is incorporated by utilising sieve plateau variance estimators based on a meaningful notion of unit relatedness.

Prioritization of therapeutic targets for cardiovascular diseases using integrative multi-omics and machine learning analyses

Lijuan Wang

Background
Treatment of cardiovascular diseases (CVD) continues to present a significant challenge, particularly for patients who are resistant to first-line CVD drugs. This study aims to identify potential repurposing opportunities for CVD treatment by integrating multi-omic data.
Methods
We applied a three-step study design. First, large-scale meta-analyses of genome wide association study (GWAS) for seven CVD related outcomes were conducted, followed by transcriptome wide association study (TWAS) and colocalization analysis. Then, the obtained differentially expressed genes (DEGs) were incorporated into step 2, in which connectivity map (CMap) analysis was performed by employing a Kolmogorov-Smirnov test. In step 3, replication analyses integrating different machine learning methods and diverse levels of data were utilized to test the validity of observed repurposed drugs and prioritize potential therapeutic targets for future drug development for CVD.

Results
The total sample size for the included CVD outcomes ranges from 30,000 to 1,000,000. As for CVD, totally 268 genomic loci, resulting in 457 lead SNPs, were identified by GWAS meta analysis. Based on the combined summary statistics, TWAS detected 224 DEGs significantly associated with CVD risk. Then, 918 compounds were identified as potential candidates that could be repurposed for CVD treatment, and 31/918 has been approved for medical use. Finally, 21 proteins were identified causally associated with CVD risk, which could be considered as potential therapeutic targets. For some known drugs, the proteins may not be their initially reported targets, indicating that the drugs could exert their anti-cardiovascular effects through multiple pathways. In addition, other medications such as antidepressants may be repurposed for CVD treatment, given their potential in responding to cellular stresses.
Conclusions
This study prioritized potential therapeutic targets for the treatment of CVD related outcomes by integrating diverse sources of data and employing advanced algorithms in a high-throughput manner, providing insights into future design of clinical trials and drug development against CVD and its subtypes.

Unveiling patterns in rare events: Statistical techniques for extreme multivariate events

Ioannis Papastathopoulos

The study of extreme events holds importance across various fields. This talk will discuss statistical models specifically designed to analyse and predict the frequency of rare, high-impact events in multivariate settings. I will present techniques based on Generalized Linear Models (GLMs) with hierarchical structures and random effects. These models are designed to appropriately connect with the underlying geometry of the distribution, leading to the ability to capture nearly any extremal dependence structure. Potential applications in fields such as anomaly detection, extreme selection and the analysis of extreme breeding values will be briefly discussed.

Multivariate genome-wide association of quantitative reading achievement scores and dyslexia diagnosis

Michelle Luciano

Individual differences in reading ability are influenced by genetic variation, with a heritability of 0.66 for word reading, estimated by twin studies. Until recently, genomic investigations were limited by modest sample size. Here we use a multivariate genome-wide association study (GWAS) method, MTAG, to leverage summary statistics from two independent GWAS efforts, boosting power for analyses of reading ability; GenLang meta-analysis of word reading (N = 27 180) and the 23andMe, Inc., study of dyslexia (Ncases = 51 800, Ncontrols = 1 087 070). We increase effective sample size of quantitative word reading to N = 102 082, with single-nucleotide polymorphism (SNP) based heritability estimated at 24%. We identified 35 independent genome-wide significant loci, including 7 regions not previously reported, but most loci represented hits found in the larger GWAS of dyslexia. Is this evidence that dyslexia, as we hypothesise, represents the low tail of the reading distribution in the population?

Challenges in integrating high dimensional biomedical data

Aris Sionakidis

The development of multimodal assays for analysing biological samples offers a significant advancement in our ability to study the complexities of biology, from the cellular level to whole organisms. These assays allow for the investigation of multiple aspects of biological diversity and complexity, shedding light on mechanisms of development, tissue organization, and the intricacies of various diseases. A fundamental challenge in the analysis of multimodal data from biological samples lies in effectively combining information across different types of data and/or different sources of the same type of data, a process often referred to as 'data integration'. This task encompasses a wide array of techniques, from correcting discrepancies between datasets (batch correction) to linking genetic variations and chromatin accessibility with gene expression patterns. High-dimensional biomedical data can be integrated horizontally, vertically or using a combination of these approaches. Although many integration strategies share underlying mathematical concepts, they are designed with different objectives in mind and are based on distinct principles and assumptions. Therefore, each integration approach comes with a set of important limitations that may limit analysis power and/or interpretability of results. As a result, there is a need for new definitions and frameworks to better understand and categorize existing methods, as well as to facilitate the development of innovative approaches to data integration in the context of complex biological systems.

Divide and Conquer Approach in Genome-wide Association Studies

Ismail Ozkaraca

Genome-wide association studies (GWAS) are vital milestone for understanding the underlying genetic basis of complex traits/diseases. At the heart of GWAS, large scale genomic data plays a pivotal role. Software tools to make use of large-scale datasets for GWAS is limited and do not evolve at the same rate as increment of data, due to cheap sequencing techniques available. In this study, we have addressed the issue of computational expense of large-scale GWAS. We have developed a method that divides a given cohort into randomly formed subcohorts, independently perform GWAS on each subcohort, and then combine the results using a novel meta-analsysis technique that considers population structure and other confounders between subcohorts. We have shown through simulations and real-data examples that our approach is effectively controlling unreliable inflation of effect sizes, a phenomenon known as winner's curse. In addition, our pipeline is suitable for incremental GWAS as per data being added and is compiled in a workflow management system, allowing users to use it in whichever compute environment they are using.
