UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

COURSE SYLLABUS

Statistical Models for Genetics

2526-1-F8205B022

Learning objectives

The course aims to equip students with the statistical tools needed to analyse and interpret genetic data, with particular reference to genome-wide association studies (GWAS). By the end of the course, students will be able to critically read a published GWAS, recognising its quality control, statistical model and replication strategy; to independently run, in the R environment, a complete pipeline that proceeds from genotype quality control to association testing, and on to population-structure correction, imputation and meta-analysis; to choose statistical models and genotype encodings appropriate to the biological question; to diagnose and correct the most common pitfalls of these studies, foremost population stratification and multiple testing; and finally to translate a statistical signal into a preliminary biological hypothesis, producing a reproducible report that documents the entire analysis.

Contents

The course covers the full analytical path of a genetic association study, organised around a single case study — a case-control GWAS for type-2 diabetes in a European population. The topics addressed include the basic vocabulary of genetic variation and data formats, allele frequencies, Hardy–Weinberg equilibrium and linkage disequilibrium; the nature of sequencing data and the critical reading of their quality; quality control at the sample and variant level; correction of population structure through principal component analysis and mixed models; single-SNP association testing with its attendant problems of effect interpretation and multiple testing; genotype imputation; meta-analysis and replication; and finally the functional annotation that leads from the statistical locus to a biological hypothesis.

Detailed program

The course opens with an introduction to genetic variation, establishing the fundamental concepts and vocabulary

(locus, allele, genotype, polymorphism, SNP, haplotype) and discussing the observational nature of human genetics, which makes statistical reasoning indispensable.

It then turns to the genetic data themselves: the formats used to represent genotypes, the calculation of allele and genotype frequencies, Hardy–Weinberg equilibrium as a reference model and its test, and linkage disequilibrium with its implications for the design and interpretation of a GWAS.

A supplementary module is devoted to how sequencing data come into being, following the chain that leads from the DNA molecule to the sequence file. The aim is not bioinformatic but statistical: to understand that the data are a noisy estimate and to learn to read a quality report critically, recognising the sources of structured noise that propagate to downstream models.

Quality control is then addressed as an integral part of every GWAS, distinguishing sample-level from variant-level filters and discussing the statistical rationale of each threshold, through to the documentation of the process.

The course then enters the theme of population structure, showing how uncorrected stratification inflates false positives and how it can be diagnosed and corrected, through principal component analysis and linear mixed models, which address stratification and cryptic relatedness in a single step.

The core of the course is single-SNP association testing: the linear and logistic regression framework, the various genotype encodings, the interpretation of effects in terms of effect size and odds ratio, the multiple-testing problem and the origin of the genome-wide threshold, and the overall diagnostics of the analysis (Manhattan plot, QQ plot, genomic inflation factor).

There follow genotype imputation, which extends coverage from the chip markers to the whole genome by exploiting linkage disequilibrium, and the use of dosages in association testing; and meta-analysis with replication, which allows multiple cohorts to be combined to gain power, their heterogeneity to be assessed, and the results to be consolidated.

The course closes with the post-GWAS phase, in which the statistical signal is connected to biology through functional annotation and public resources, and with a guided project on a second dataset, in which students replicate the entire pipeline and produce a clinical-style report.

Prerequisites

A solid grounding in classical biostatistics is required: statistical inference, hypothesis testing, confidence intervals, and familiarity with linear and logistic regression models. A basic knowledge of the R language and the RStudio environment is necessary, sufficient to manipulate data and produce plots. No prior background in genetics or bioinformatics is required: the necessary biological concepts are introduced during the course to the extent useful for the statistical analysis.

Teaching methods

The course alternates lectures with hands-on R laboratories, run on students' own laptops in the RStudio environment, with no need for dedicated computing infrastructure. The teaching design is driven by a single case study followed from beginning to end: each new method is introduced starting from the clinical or epidemiological question that motivates it, developed as a statistical model, and finally translated into code. The laboratories are not isolated exercises but progressively build the successive segments of a single pipeline, which converges into the final deliverable. Guided reading of scientific papers accompanies the lectures, training students to critically

interpret the methods, statistics and figures of a published GWAS.

Assessment methods

Assessment is based on a final deliverable consisting of a reproducible R Markdown report, built on a second case-control dataset different from the one used in class and provided by the instructor. The examination is structured in three interconnected components: a guided project, carried out in a supervised laboratory, in which the student replicates the entire pipeline and gives shape to the deliverable; submission of the report, subject to the requirement that it can be re-run in full on the instructor's laptop, from raw data to tables and figures, without manual intervention; and an oral discussion of about fifteen minutes, individual or in small groups, in which the methodological choices and results are presented and critically defended.

The deliverable and its discussion are assessed against four criteria of equal weight, each accounting for twenty-five per cent of the overall judgement. The first is methodological correctness, that is, the appropriateness of each step of the pipeline relative to the question and the data, with explicit justification of the choices of encoding, covariates and thresholds. The second is statistical interpretation, namely the ability to read results in terms of effect size, confidence and uncertainty, and not merely of p-values. The third is clinical and biological interpretation, that is, placing the top loci in their biological context using public resources, with due caution about the locus/gene distinction. The fourth is reproducibility, understood as the report's ability to be re-run end-to-end without manual intervention.

The grade is expressed on the Italian thirty-point scale. A pass, set at eighteen out of thirty, requires a pipeline that is correct in its essential steps and a report that re-runs without errors; a full mark additionally requires accurate statistical and clinical interpretation and fully justified methodological choices. The distinction "cum laude" is reserved for deliverables that demonstrate critical independence, for example through sensitivity analyses that were not requested, discussion of violated assumptions, or a particularly insightful use of annotation resources. Reproducibility constitutes a requirement for admission to assessment: a report that does not run is returned for correction before being assessed on its merits. Small-group work is permitted for the project, but the oral discussion verifies individual contribution and understanding. Attendance at the laboratories, while not graded in itself, is strongly recommended, since each of them builds a part of the pipeline that converges into the deliverable.

Textbooks and Reading Materials

For a concise overview of the entire GWAS pipeline, see Bush and Moore, Genome-Wide Association Studies (PLOS Computational Biology, 2012), and the applied tutorial by Marees and colleagues, A tutorial on conducting genome-wide association studies (Int J Methods Psychiatr Res, 2018). For thematic depth, the following are recommended: Price and colleagues, New approaches to population stratification in genome-wide association studies (Nature Reviews Genetics, 2010); Marchini and Howie, Genotype imputation for genome-wide association studies (Nature Reviews Genetics, 2010); and Evangelou and Ioannidis, Meta-analysis methods for genome-wide association studies and beyond (Nature Reviews Genetics, 2013). As a concluding synthesis, Visscher and colleagues, 10 years of GWAS discovery: biology, function, and translation (American Journal of Human Genetics, 2017), is particularly useful. For the module on sequencing data, see Goodwin, McPherson and McCombie, Coming of age: ten years of next-generation sequencing technologies (Nature Reviews Genetics, 2016), and, on the origin of the quality score, Ewing and Green, Base-calling of automated sequencer traces using phred. II (Genome Research, 1998).

Semester

Second semester

Teaching language

italian

Sustainable Development Goals

GOOD HEALTH AND WELL-BEING