Study Guide

Bioinformatics and Statistical Genomics (SGX)

Semester 2, 2021

Prepared by:
Prof David Balding

Melbourne Integrative Genomics,
School of BioSciences and School of Mathematics & Statistics,
University of Melbourne

Copyright © University of Melbourne

Table of Contents

1. Instructor details

Prof David Balding

Melbourne Integrative Genomics, Building 184, Royal Parade, the University of Melbourne, VIC 3010 Ph: (03) 3844 3730 internal 43730

Email: dbalding@unimelb.edu.au

The course will be delivered by the co-ordinator and Dr Sudaraka Mallawaarachchi, also based at MIG. Email: sudaraka.mallawaarachchi@unimelb.edu.au

2. Overview

Statistical genomics is the application of statistical methods to understand genomes, their structure, function and evolutionary history, in many different scientific contexts, including: understanding biological mechanisms in health and disease, optimising economic and welfare traits in animals and plants, learning about the history of humans and other organisms, and identifying individuals and their relatedness. **Bioinformatics** is an overlapping term that suggests more emphasis on data management and software pipelines. **Genetic epidemiology** is another closely related field, in which statistical genomics methods are used with family or population data to study causes of disease.

Statistical genomics is characterised by large datasets, high-dimensional regression models, stochastic processes, and computationally-intensive statistical methods. In this unit we will learn about

- some of the relevant biology and terminology,
- important mathematical models and inference methods in medical, population and evolutionary genetics,
- how to test for association between genetic variants and outcomes of interest,
- genome-wide statistical models to understand the genetic mechanisms underlying a trait and to predict outcomes,
- sequence analysis using hidden Markov models,
- an introduction to some other analyses “downstream” from the genome, including gene expression and epigenetics, and microbiome analysis.

The statistical package R will be employed to perform analyses.

3. Unit Objectives

At the end of this unit you should be able to:

1. Describe core mechanisms of genetics, including mutation, recombination and selection.
2. Use the Wright-Fisher and coalescent models of population genetics for simulation and inference.
3. Perform sequence analysis using hidden Markov models.
4. Describe the key data, models and inference goals of phylogenetics.
5. Access genomic data from public databases.
6. Perform a genetic association analysis, including the assessment of possible confounding.
7. Explain the concept of heritability and its estimation.
8. Use genome-wide SNP data to develop prediction models.
9. Explain key features of data and statistical models used in the fields of transcriptomics, epigenetics, and bacterial genomics.
10. Effectively communicate results of statistical analyses in genomics and related areas.

4. Unit content

The unit is divided into eight modules, with either one or two weeks devoted to each. These are listed below. The “weeks” column is indicative and is provided as a guideline.

Module	Weeks	Content
1	1	Basics of human genetics and genetic epidemiology, review of R
2	2,3	Population genetics, neutral Wright-Fisher and coalescent models
3	4,5	Genetic association analysis including GWAS
4	6	Introduction to transcriptomics, epigenetics and bacterial genomics
5	7	Heritability and genomic prediction
6	8,9	Sequence analysis using hidden Markov models
7	10	Genomic medicine
8	11,12	Evolutionary models, selection and phylogenetics

The course is designed for a student time commitment of up to 12 hours per week, of which on average 2 hours per week will be devoted to assessed assignments (= 8 hrs per assignment), 2 hrs to self-assessed exercises and online discussions, and up to 8 hours per week studying the course materials.

5. Method of Delivery and Communication

The unit is offered in distance mode. Access to course notes, data sets, exercises and solutions, as well as submission of assignments, will be via Canvas. We will also make use of Zoom meetings through Canvas and videos via the Studio link.

Communication from instructors to all students will be via announcements in Canvas. During semester it will be assumed that you have read announcements on Canvas within 48 hours of posting (excluding weekends). Direct, private communication to individual students will be through the inbox email facility in Canvas.

Students wishing to communicate with instructors should use the discussion forum in Canvas for any academic matter. Please use the the Canvas inbox for direct, private messages concerning any personal issues, such as health or other factors affecting your progress.

Student "Profiles"

In order to introduce ourselves to each other, we encourage you to populate your Canvas profile with some information about yourself, such as any personal interests you are happy share, and

also practical things like your preferred name and contact details, your academic and professional background, your job role and employer, and what you hope to gain from studying SGX.

Feedback

Your constructive comments are valued and guide us in the continuous improvement of the unit.

6. Assessment and semester activity

There are 12 teaching weeks from Monday July 26 until Friday October 22. The mid-semester break is the week of 27 Sep - 1 Oct

Assessment for the unit consists of: three assignments during the semester teaching period (20% + 20% + 20% = 60%) and a final at-home written examination (40%) over 4 days during the exam period (end Oct/early Nov).

Assignments

Rules and instructions for assignment submission are given in the document Assessment Guide which is available on the BCA website.

The assignments will be set:

Assignment 1: Friday of week 2 (Aug 6)

Assignment 2: Friday of week 6 (Sept 3)

Assignment 3: Friday of week 10 (Oct 8)

Due dates will usually be the 2nd Tuesday after the assignment is set (11 days later). This will be confirmed when the assignment is set.

***Late submission*:** Unless otherwise stated, a student can submit an assessment up to 5 days after the due date. A late penalty of will be applied which is to multiply the final mark by 0.95 for each day late (including weekends and public holidays). However, no penalty will be applied to marks below 50% and if the late penalty would have reduced the mark below 50% then exactly 50% will be awarded.

Final Assignment/Examination

The examination will be released on Canvas by **5pm on Friday October 29, and is due at 5pm on Tuesday November 2** (so you have two weekend days and two weekdays to complete the exam, which should require about 4-6 hours work). You should not communicate with other students, or any other person, during the exam period in any topic connected with the examination.

Exercises

There are self-assessment exercises in the notes, in addition to the assessed assignments. Online discussions about these exercises is welcome, but not about the assessed assignments or exam.

Online participation

You are strongly encouraged to participate in online discussions which can help create a group ethos. Students are encouraged to try to answer each other's questions, which will be moderated by the Instructors in case a reply is incorrect. Discussion of topics that form part of a current assessment task is permitted after the due date.

7. Textbooks

Handbook of Statistical Genomics (Eds: Balding, Marioni and Moltke, 4th ed, Wiley 2019).

which has 36 chapters summarising the start-of-art in the field, as well as an extensive glossary. The *Handbook* is available online through your university library, please check as soon as possible that you can access it. Some libraries may also have a print version available – the University of Melbourne library has both. The *Handbook* is huge at well over 1,000 pages and we will only examine a small fraction of it in this course, but you should take the opportunity to browse other chapters/sections to get a fuller understanding of the field.

You also have access through your university library to 18 online lectures on Statistical Genetics offered by **Henry Stewart Talks** in their **Biomedical & Life Sciences Collection**. Access details may vary according to your home university so please check as soon as possible: you should be able to access directly **here**, if there is a problem please try through your library's online catalogue.

Other useful texts:

- The Fundamentals of Modern Statistical Genetics, Nan Laird, Christoph Lange, Springer, 2011
- Applied Statistical Genetics with R: For Population-based Association Studies, Andrea S. Foulkes Springer 2009.
- W Ewens, G Grant. *Statistical methods in bioinformatics - an introduction*, Springer 2005.
- R Durbin, S Eddy, A Krogh, G Mitchison. *Biological Sequence Analysis*, Cambridge UP 1998.

8. Software

We will be using the statistical package R. You can download and install the latest version of this reliable freeware from the R homepage. You should also install Rstudio, a free graphical user interface for R which has many advantages for users.

9. Useful websites

- BCA Canvas homepage
- BCA homepage
- BCA Student Resources (including the Guide for Reporting Statistical Results)

10. Complaints policy

Please see the BCA complaints policy in the Assessment Guide