BIO
STATISTICS
COLLABORATION
OF AUSTRALIA

Study Guide

Bioinformatics and Statistical Genomics (SGX)

Semester 2, 2023

Prepared by:

Associate Professor Agus Salim

Melbourne School of Population and Global Health and
School of Mathematics & Statistics,
University of Melbourne

THE UNIVERSITY OF
MELBOURNE

Contents

Bioinformatics and Statistical Genomics (SGX)

Semester 2, 2023

Contact details

Associate Professor Agus Salim

Subject Coordinator and Instructor

Melbourne School of Population and
Global Health and School of Mathematics
and Statistics, The University of
Melbourne

(03) 8344 7953

salim.a@unimelb.edu.au

If you have any general BCA queries, please contact: Karolina Kulczynska-Le Breton or Emily Higginson at the BCA Coordinating Office on 02 9562 5076/54 or email bca@sydney.edu.au

Background

The advance in genomics sequencing technology in recent years have increased the feasibility of conducting large scale, population-level genetic studies. These studies are wide-ranging in their purposes, from understanding disease aetiology to genomic-based risk prediction with personalised medicine being the ultimate objective. Biostatisticians working in health and medical research and related industries increasingly required to handle genomics data and prior experience with genomics data is advantage in such situations.

Context within the program

This unit is an elective subject offered in Semester 2 every second year. The unit draws prior knowledge learnt in theory-oriented subjects (MBB, PDT, PSI) and applied-oriented subjects (DMC, LMR) to introduce students to several statistical methods that have been developed to tackle problems in statistical genomics and bioinformatics.

Prerequisites
Mathematical Background for Biostatistics (MBB)
Probability and Distribution Theory (PDT)
Principles of Statistical Inference (PSI)
Data Management & Statistical Computing (DMC)
Linear Models (LMR)

Co-requisites
None

Unit summary

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.

Workload requirements

The expected workload for this unit is 10-12 hours per week on average, consisting of guided readings, discussion posts, independent study, and completion of assessment tasks.

Learning Outcomes

At the completion of this unit students 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.

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. Each module will have the following materials:

1. Module notes containing description of concepts and methods and links to relevant online lectures, published articles and textbooks.
2. Self-assessment exercises.

Study materials for all Modules are accessed from the eLearning unit site. Assignments and supplementary material such as datasets will be available within each Assignment item. Please note that we are not able to post copies of copyright material (journal articles and book extracts)—for these you will have to rely on your home university's library.

Recommended approaches to study

Students should work through each module systematically, following the module notes and any readings referred to, and working through the accompanying exercises. *You will learn a lot more efficiently if you tackle the exercises systematically as you work through the notes.* You should also work through all the computational examples in the notes for yourself on your own computer.

Outline solutions to the exercises in each module (except those to be submitted for assessment, as described below) will be posted online at the midway point of the allocated time for the module. This is intended to encourage you to attack the exercises independently (or via the eLearning site), and yet not make you wait too long to see the sketch solutions.

Make the most of this unit by engaging with coordinators and fellow students on the Discussion Board and in Tutorials. These are safe spaces to discuss the course material and related ideas and students are encouraged to make the most of them by engaging in respectful discussion.

Questions about Assignments should be directed to the coordinator in the first instance to avoid any Academic Honestly issues.

Method of communication with coordinator(s)

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.

Questions about administrative aspects or course content can be emailed to the coordinator. Please use "(BCA code):" in the Subject line of your email to assist in keeping track of our email messages. Coordinator/s will be available to answer questions related to the module notes and practical exercises, and to address any other issues that require clarification.

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 Board in Canvas for any academic matter. Please use the Canvas inbox for direct, private messages concerning any personal issues, such as health or other factors affecting your progress.

Please note that instructors are not necessarily available every day of the week and you should expect that it may take a day or so to respond to questions (possibly longer over weekends and during breaks). We strongly recommend that you post content-related questions to the Discussion Board in the unit site.

Module descriptions

Below is an outline of the study modules, followed by a timetable and assessment description table

Module 1: Basics of human genetics and genetic epidemiology, review of R

Topic

- Introduction to Human Genetics
- Basic Statistics for Genetics
- Overview of Genetic Epidemiology

Module 2: Population genetics, neutral Wright-Fisher and coalescent models

Topic

- Introduction to Population Genetics
- Coalescent Models

Module 3: Genetic association analysis including GWAS

Topic

- Introduction to Genetic Association Analysis
- Imputation in GWAS
- Intro to GWAS replication and meta analysis

Module 4: Introduction to transcriptomics, epigenetics and bacterial genomics

Topic

- Transcriptomics
- Epigenetics
- Microbiome analysis and Bacterial Genomics

Module 5: Heritability and genomic prediction

Topic

- Traditional heritability
- SNP-based heritability
- Genomic prediction

Module 6: Sequence analysis using hidden Markov models

Topic

- Intro to hidden Markov models
- Modelling protein families using HMM

Module 7: Genomic Medicine

Topic

- Variant interpretation and Genomic Medicine
- Intro to databases and software in Genomic Medicine

Module 8: Evolutionary models, selection and phylogenetics

Topic

- Models of nucleotide and amino acids evolution
- Selection
- Phylogenetics

Unit schedule

Semester 2, 2023 starts on Monday July 24, 2023

Week	Week commencing	Module	Topic	Assessment
1	Jul 24, 2023	Module 1		
2	Jul 31, 2023	Module 2		Assignment 1 available August 4
3	Aug 7, 2023	Module 2		
4	Aug 14, 2023	Module 3		assignment 1 due August 15
5	Aug 21, 2023	Module 3		
6	Aug 28, 2023	Module 4		Assignment 2 available Sept 1
7	Sept 4, 2023	Module 5		
8	Sept 11, 2023	Module 6		assignment 2 due Sept 12
9	Sept 18, 2023	Module 6		
	Sept 25, 2023		**mid semester break** 1 week only	
10	Oct 2, 2023	Module 7		Assignment 3 available Oct 6
11	Oct 9, 2023	Module 8		
12	Oct 16, 2023	Module 8		assignment 3 due Oct 17
13	Oct 23, 2023	Exam Prep		Exam available @5pm on Oct 27 (Fri) and due @5pm on Oct 31 (Tue)

Assessment

Assessment includes 3 written assignments worth 20% each. Each assignment will be made available on a specific date (see Table above), and to be completed within 11 days. Unless otherwise stated, assessments are due by 11:59pm on the stated day.

At the end of the semester, there will be a take-home exam (worth 40%) to completed over 4 days (see Table above for exam schedule).

Assessment name	Assessment type	Coverage	Learning objectives	Weight
Assignment 1	Assignment	Modules 1,2	1,2	20%
Assignment 2	Assignment	Module 3,4	1,6,9,10	20%
Assignment 3	Assignment	Module 5-6	1,3,7,10	20%
Take-home Exam	Exam	Modules 1-8	1,2,3,4,5,6,7,8,9,10	40%

In general, you are required to submit work typed in Word or similar. We strongly recommend you become familiar with equation typesetting software such as Microsoft's Equation Editor for algebraic work. You may submit neatly handwritten work, however please note that marks will potentially be lost if the solution cannot be understood by the markers due to unclear or illegible writing. Handwritten work should be scanned and collated into a single pdf file and submitted via the eLearning site. See the BCA Assessment Guide for guidelines on acceptable standards for assessable work.

Students are encouraged to discuss relevant topics in the Discussion Board. However, please avoid posting questions relating directly to assessable material. These should be emailed to the Unit Coordinator in the first instance.

Explicit solutions to assessable exercises should not be posted for others to use. Each student's submitted work must be clearly their own, with anything derived from other students' discussion contributions clearly attributed to the source.

Submission and academic honesty policy

All assessment material should be submitted via the relevant Assessment module in Canvas unless otherwise advised. Turnitin plagiarism detection is applied to all submissions. For detailed information, please see the BCA Assessment Guide, which includes links to the Academic Honesty policies at member universities. Please familiarise yourself with the procedures and policies at your home university. You will need to indicate your compliance with the plagiarism guidelines and policy at your home university.

A special note regarding "contract cheating" sites: Unfortunately, there have been instances in the past of students using such websites to post assignment questions and receive solutions (usually for a fee). We have arrangements with these sites to identify the student posting questions or accessing the solutions, and such students will be referred to and face disciplinary processes at their home university.

Use of ChatGPT or other generative AI tools in assessment tasks

The assessment tasks in this Unit have been designed to be challenging, authentic and complex. Although individual assessment components may provide specific guidance regarding the use of generative AI tools (e.g., ChatGPT), successful completion of these components will require students to critically engage in specific contexts and tasks for which artificial intelligence will provide only limited support and guidance. In all cases, a failure to reference the use of generative AI may constitute student misconduct under the Student Code of Conduct of your University of enrolment. To successfully complete assessment tasks, students will be required to demonstrate detailed comprehension of their written submission independent of AI tools.

Late submission and extension procedure

The standard BCA policy for late penalties for submitted work is a 5% deduction from the earned mark for each day the assessment is late, up to a maximum of 10 days (including weekends and public holidays). Extensions are possible, but these need to be applied for (by email) as early as possible. The Unit Coordinator can approve extensions up to three days; for extensions beyond three days, you must apply to your home university, using their standard procedures.

Learning resources

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.

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.

Software requirements and assumed knowledge

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.

For help with R, please see Learning R in the Student Resources site.

If you have not yet organised access to these packages, you should do so as soon as possible. This is a practical course which requires regular use of the relevant software; delays in gaining access to these packages may impact your ability to complete the course. Information on how to download R and RStudio can be found in the BCA Textbook and Software Guide.

Required mathematical background

Familiarity with statistical concepts and models covered in MBB, PDT, PSI and LMR.

Feedback

Our feedback to you:

The types of feedback you can expect to receive in this unit are:

- Formal individual feedback on submitted exercises assignments
- Responses to questions posted on Canvas

Your feedback to us:

One of the formal ways students provide feedback on teaching and their learning experience is through the BCA student evaluation survey at the end of each semester. The feedback is anonymous and provides the BCA with evidence of aspects that students are satisfied with and areas for improvement.

Unit changes, including response to recent student evaluation

SGX was last delivered in Semester 2 2021. Based on students feedback in 2021, Online Genetic Dictionary/Encyclopedia has been added as resources on Canvas Site to help students understand the concepts introduced in this subject better.

Acknowledgments

The study coordinator would like to acknowledge David Balding and Sudaraka Mallawaarachchi (both of the University of Melbourne) who prepared the original course materials.