

Course Overview & Short Introduction

Valeria Vitelli

Oslo Centre for Biostatistics and Epidemiology

Department of Biostatistics, UiO

valeria.vitelli@medisin.uio.no

MED3007

Statistical Principles in Genomics: an Introduction with Rstudio

13.01.2025

- 1 Course Overview
 - Course summary & schedule
 - Take-home Exam

- 2 Short Introduction to the Topic
 - Common statistical tasks with large-scale biological data
 - Group Work

Course Content

Motivations

- molecular (genomic) data increasingly important in medical research & clinical practice
- such data are *very high-dimensional* → challenges

Aims

- **general:** learn the challenges & solutions when analysing genomic data, both in **theory** and **practice**
- **theory:** focus on three main tasks; screening, visualisation, clustering
- **practice:** introduction to RStudio, many examples

Course Structure

General structure

- afternoon:
 - lecture on a topic + Q & A session
 - some free time for reflection / group work
- morning after: practical lab session on the same topic

	Monday	Tuesday	Wednesday	Thursday	Friday
9:00 - 9:45	Intro Course	Intro Lab 1	Lab 2	Lab 3	Exam
10:00 - 11:30	Intro RStudio	Lab 1	Exercises	Exercises	
11:30 - 12:30	lunch	lunch	lunch	lunch	
12:30 - 13:30	Lecture 1	Lecture 2	Lecture 3	Exam Sim	
13:30 - 15:30	Group Work	Group Work	Group Work	Wrap-up	

Special Sessions

Thursday is quite interactive. Used to fix concepts from the course, to ask questions, and to see a take-home exam simulation.

Practical Info

Website

More interactive website we will use for teaching materials:

https://ocbe-uio.github.io/course_med3007/

Canvas

- Course webpage on Canvas:
<https://www.uio.no/studier/emner/medisin/med/MED3007>
- <https://uio.instructure.com> and Mobile App Canvas Student (for iOS and Android)
- All course material (lecture notes, computer labs, data sets, reading material) will be ALSO made available via Canvas
- If you have problems logging in, send me email asap!

Info on the course Take-home Exam

Idea

- The purpose of the exam is to **replicate** one of the analyses we will see in class
- **Any proposal will be accepted.** The focus is not on right/wrong answers, but on showing independence in managing high-dimensional data analysis: choosing a method for a given purpose

Details

- Starting point is a **data set**, which will be provided via **Inspira** on **Thursday at 15:00**
- You have time until **Monday** same time
- You can provide your analysis as a .pdf or .doc file, or as a .R (or .Rmd) file. Important is to report what you think is relevant for me to understand *what* you did and *why* you did what you did

- 1 Course Overview
 - Course summary & schedule
 - Take-home Exam

- 2 Short Introduction to the Topic
 - Common statistical tasks with large-scale biological data
 - Group Work

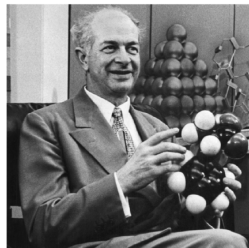
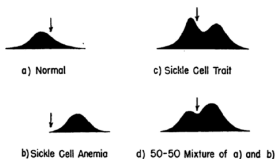
Sickle cell anemia, a molecular disease (Pauling et al, 1949)

November 25, 1949, Vol. 110

SCIENCE

543

Sickle Cell Anemia, a Molecular Disease¹

Linus Pauling, Harvey A. Itano,² S. J. Singer,² and Ibert C. Wells³*Gates and Crellin Laboratories of Chemistry,
California Institute of Technology, Pasadena, California⁴*

Sickle cell anemia, a molecular disease (Pauling et al, 1949)

- Pauling et al. showed that hemoglobin from patients suffering from sickle cell anemia had a different electrical charge than that from healthy individuals.
- This report had a powerful impact on both the biomedical community and the general public for two reasons:
 - 1 It showed that the cause of a disease could be traced to an alteration in the molecular structure of a protein.
 - 2 As this disease was known to be inherited, the paper argued that genes determine the structure of proteins.

Biomarkers / molecular markers

Biomarker

We refer to a biomarker as a biological quantitative measure associated with a clinical outcome. It may be a single trait, or a grouping (signature) of traits that separates different populations with respect to an outcome of interest.” (Sargent *et al.*, 2005)^a.

Examples: classic laboratory parameters (blood pressure, cholesterol level, blood glucose) and molecular markers.

^aSargent, D. J., Conley, B. A., Allegra, C., & Collette, L. (2005). Clinical trial designs for predictive marker validation in cancer treatment trials. *Journal of Clinical Oncology*, 23(9), 2020-2027.

Molecular marker

A molecular marker is a biomarker that can be detected using genomics and/or proteomics technologies.

Examples: gene mutations, gene expression, protein expression, methylation patterns.

Molecular diagnostics

It means **to identify molecular markers in the genome by applying molecular biology to medical testing**. The technique is used to diagnose and monitor disease, detect risk, and decide which therapies will work best for individuals.

Examples:

- **Prenatal tests** for chromosomal abnormalities (e.g. Trisomy 21)
- **Infectious diseases:** test for infectious diseases such as chlamydia, influenza virus and tuberculosis; or for specific strains such as the **H1N1 virus** (pathogenomics).
- **Cancer screening:** for example testing for mutations in BRCA1 or BRCA2 for familial form of breast cancer (about 5% of all breast & ovarian cancers) → “Angelina Jolie” effect

Statistical Principles in Genomics

- The main difference to “classical statistics” is the **high-dimensionality of the data**: in genomic data we have tens of thousands or millions of input variables/ features.
- Most of those features will not be of interest in the specific context of the experiment. → The main task is the **identification of important features** (genes, SNP's, etc).
- We have to deal with the curse of dimensionality.

References:

- 1 James et al. (2013) An Introduction to Statistical Learning with Applications in R, <https://www.statlearning.com/>, Chapters: 1, 2, 12, 13
- 2 Holmes, Huber (2019). Modern Statistics for Modern Biology, <http://web.stanford.edu/class/bios221/book/> Chapters: 3, 5, 6, 7

Statistical Principles in Genomics

- ① **Screening and multiple testing**
(to determine lists of differentially expressed genes)
- ② **Exploratory analysis** and **unsupervised learning**
 - **Exploratory analysis: Dimension reduction and visualisation**
(ex. principal components analysis)
 - **Unsupervised learning: Clustering and heatmaps**
(ex. to find subgroups of similarly regulated genes)

	Monday	Tuesday	Wednesday	Thursday	Friday
9:00 - 9:45	Intro Course	Intro Lab 1	Lab 2	Lab 3	Exam
10:00 - 11:30	Intro RStudio	Lab 1	Exercises	Exercises	
11:30 - 12:30	lunch	lunch	lunch	lunch	
12:30 - 13:30	Lecture 1	Lecture 2	Lecture 3	Exam Sim	
13:30 - 15:30	Group Work	Group Work	Group Work	Wrap-up	

Info on the course Group Work

Idea

- The purpose of the group work is to **find practical examples** of the analyses we will see in class
- Aim: **Get the discussion going.** The focus is not on right/wrong answers, but on discussing issues seen in class in a practical setting

Details

- Starting point: **two papers**, paper 1 and paper 2 (see next slides)
- Paper 1 is related to Lecture 1 and 2; Paper 2 to Lecture 3
- **Group Work: what are you supposed to do?** Read the paper, and try to answer in groups to the following questions
 - 1 Which was the main research question addressed in the paper?
 - 2 Where do results of methods seen in class are reported in the paper? (which figures/tables)

Paper 1 (focus for Lectures 1 & 2)

Cappelletti et al. *Clinical Proteomics* (2022) 19:23
<https://doi.org/10.1186/s12014-022-09361-1>


Clinical Proteomics

RESEARCH

Open Access



Quantitative proteomics reveals protein dysregulation during T cell activation in multiple sclerosis patients compared to healthy controls









Chiara Cappelletti¹, Anna Eriksson³, Ina Skaara Brorson^{3,4}, Ingvild S. Leikfoss^{3,4}, Oda Kråbøl², Einar August Høgestøl^{3,4,5}, Valeria Vitelli⁶, Olav Mjaavatten⁷, Hanne F. Harbo^{3,4}, Frode Berven⁷, Steffan D. Bos⁴ and Tone Berge^{1,2*} 

Paper 2 (focus for Lecture 3)

Published online 24 March 2022

NAR Cancer, 2022, Vol. 4, No. 1 1
<https://doi.org/10.1093/narcan/zcac008>

Epigenetic alterations at distal enhancers are linked to proliferation in human breast cancer

Jørgen Ankill ^{1,2}, Miriam Ragle Aure ³, Sunniva Bjørklund ³, Severin Langberg ⁴, Oslo Breast Cancer Consortium (OSBREAC), Vessela N. Kristensen ³, Valeria Vitelli ⁵, Xavier Tekpli ³ and Thomas Fleischer ^{1,*}

¹Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital, Oslo, Norway, ²Institute of Clinical Medicine, Faculty of Medicine, University of Oslo, Oslo, Norway, ³Department of Medical Genetics, Institute of Clinical Medicine, Faculty of Medicine, University of Oslo, Oslo, Norway, ⁴Cancer Registry of Norway, Oslo, Norway and ⁵Oslo Centre for Biostatistics and Epidemiology, Department of Biostatistics, Faculty of Medicine, University of Oslo, Oslo, Norway