2020 - 2021

Bioinformatics

UNIVERSITY OF CAMBRIDGE

NST Part II BBS minor (Bioinformatics)
NST Part II BBS (Zoology)
NST Part II PDN
NST Part II BBS (PDN)

Bioinformatics is an interdisciplinary field that uses computational approaches to process biological data. With the biological and biomedical sciences becoming more data-driven than ever before, bioinformatics is central to these areas. The NST Part II BBS Bioinformatics minor subject introduces the fundamental bioinformatic concepts and methodologies used to analyse biological data. It is structured around 3 main blocks; data science, omics analysis, and network analysis.

The course is specially designed for students coming from the biological and biomedical sciences. The course consists of a set of lectures that introduce theoretical concepts, and practicals which provide hands-on practice using real biological datasets.

This year due to the COVID-19 pandemic, the course will be taught online via an online training environment that we have in place. We aim to provide a classroom experience as closely as possible, with opportunities for one-to-one discussion with the course lecturers and supervisors.

We hope that you and your loved ones stay safe and we look forward to welcoming you to the course in January 2021.

Alexia Cardona
Course Organiser
NST Part II BBS Bioinformatics
Email: ac812@cam.ac.uk

Cathy Hemmings
Course Administrator
NST Part II BBS Bioinformatics
Email: cgh32@cam.ac.uk

For further information on the course, please contact either the Course Organiser, Dr Alexia Cardona or the Course Administrator Cathy Hemmings .

NST II BBS Bioinformatics minor

Learning Aims

The course has the following aims:

1. To introduce students to popular methods used in Bioinformatics and their application to biological data. This would help them understand and interpret the application of these methods to cutting-edge omics and other bioinformatic analysis.
2. To develop students' skills in foundational data science methods and provide computational, statistical and machine learning knowledge required to analyse biological data and systems.
3. To provide an introduction to omics applications, focusing on the use of next generation sequencing, RNA-seq and Genome Wide Association Studies (GWASs).
4. To introduce gene ontologies and gene-set enrichment analysis used to link downstream analysis results back to biology and introduce basic concepts of biological networks and their analysis.

Learning outcomes

At the end of the course, students would:

1. Write and run scripts and apply statistical and machine learning methods to analyse biological data.
2. Understand the different stages involved in processing and analysing omics data. Perform basic variant calling and RNA-seq data analysis and be familiar with open-source software packages used to perform such analysis.
3. Explain basic concepts and be aware of methods used in biological networks and their analysis and use gene ontologies and gene-set enrichment analysis methods to map results of bioinformatic analysis back to their biological function.

After attending this module, students will not be independent in the analysis of complex biological data but will have acquired the critical thinking needed to understand what the analysis of genomic data entails, what are the strengths and weaknesses of different analysis strategies, and will be equipped with a basic set of bioinformatics skills that will enable them to explore and interpret genomic data, as well as other types of biological data, available in the public domain.

Teaching

The course consists of a set of lectures that introduce theoretical concepts, and practicals which provide hands-on practice using real biological datasets. Q&A sessions or supervisions will be provided for the different course blocks to provide support to the students and give them the opportunity to interact with the tutors and discuss questions they might have.

This year the course will be delivered online. Practicals will be delivered via our online training environment. Students will be given access before the course. As such there will be no need to install software in advance. Software installation instructions will be provided. In addition, IT support will also be provided in case students would like to install the software on their computer. We aim to provide a classroom experience as close as possible, with opportunities for one-to-one discussion with tutors.

Examination

Assessment will be via a 3-hour written exam paper. More details can be found on the course VLE.

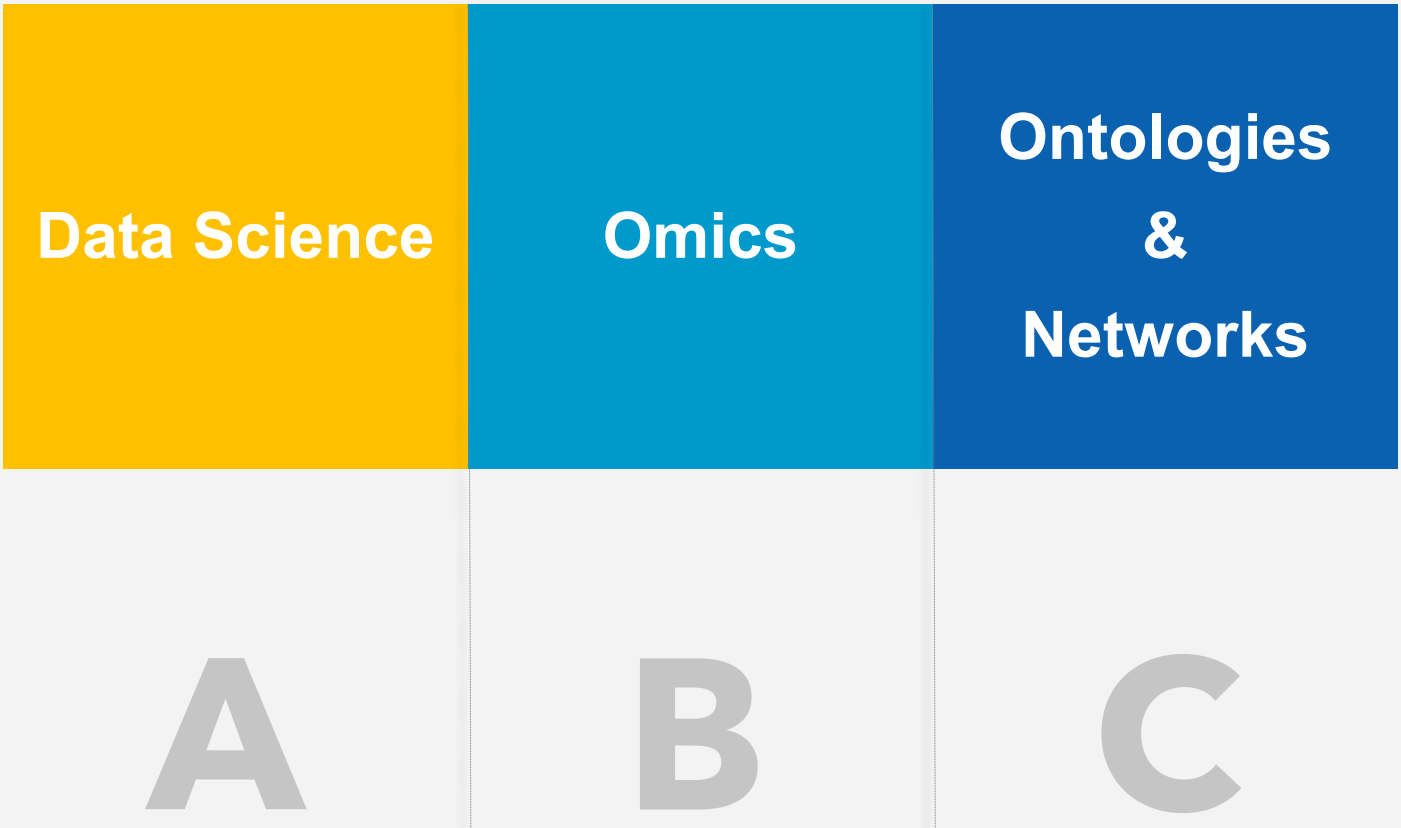
Course structure

The course starts with a key opening lecture that will first discuss the distinction between bioinformatics and computational biology, then a brief history of bioinformatics from the perspective of the growth of genomics and genome sequence is provided. Following is a discussion with examples of three core strands that feed into modern bioinformatics: algorithms, statistics and databases, and conclude with a few thoughts about likely future directions.

Opening Lecture

Title	Lecturer	Day	Date	Time	Format
Introduction to Bioinformatics and Computational Biology	Durbin	Tue	19 Jan 2021	16:00-17:00	L

The rest of the course is structured around 3 main blocks; Data Science for Bioinformatics (Block A), Bioinformatics Approaches to Omics Analysis (Block B), and Pathway and Network Analysis of Biological Data (Block C).



Block A

Data Science for Bioinformatics

01	02	03	04
Command Line		Statistics	Machine Learning

In the Data Science for Bioinformatics block, we will introduce programming, data visualisation, data manipulation, statistics and machine learning that are popular in the bioinformatics field. These topics are fundamental to the analysis of data which are currently in high-demand due to the data-driven approach of answering research questions, driven by the increasing amount of data becoming available. Knowledge gained from this set of lectures and practicals can be applied and transferred to different research domains.

01 Command Line

Title	Lecturer	Day	Date	Time	Format
Introduction to the Unix Command Line	Martinez Cuesta	Mon	18 Jan 2021	15:00-17:00	P

Using the Linux operating system and the bash command line interface, we will demonstrate the basic structure of the UNIX operating system and how we can interact with it using a basic set of commands. Applying this, we will learn how to navigate the filesystem, manipulate text-based data and structure simple pipelines out of these commands.

02 R

Title	Lecturer	Day	Date	Time	Format
Introduction to programming in R	Cardona	Mon	25 Jan 2021	15:00-17:00	P
Introduction to data visualisation and manipulation in R	Cardona	Mon	08 Feb 2021	15:00-17:00	P

R is a popular programming language in data science. In the first practical, we will learn the basic programming concepts. We will then learn about the package tidyverse and how we can use it to manipulate and visualise our data effectively.

03 Statistics

Title	Lecturer	Day	Date	Time	Format
Statistics for Large Datasets –	Castle	Thu	04 Feb 2021	16:00-17:00	L
Introduction to Statistics—Linear Regression	Mohorianu	Tue	09 Feb 2021	16:00-17:00	L
Introduction to Statistics—Linear Regression	Mohorianu	Mon	15 Feb 2021	15:00-17:00	P

Statistics is an important component to the analysis of data. We will learn about statistics methods that are used in population data and large datasets. We will understand why there is a multiple comparison problem in a range of bioinformatics techniques and what correction techniques are available. We will then learn about Linear Regression, an approach for modelling the relationship between a scalar response and one or more explanatory variables (simple linear regression and multiple linear regression, respectively). In this lecture we will focus on approaches for estimating coefficient, assessing their accuracy, and underlying some limitations of these methods, in the context of big data analyses.

04 Machine Learning

Title	Lecturer	Day	Date	Time	Format
Introduction to Machine Learning—Unsupervised Learning	Mohorianu	Tue	23 Feb 2021	16:00-17:00	L
Introduction to Machine Learning—Supervised Learning	Mohorianu	Thu	25 Feb 2021	16:00-17:00	L
Introduction to Machine Learning—Examples of Unsupervised Learning.	Mohorianu	Tue	02 Mar 2021	16:00-17:00	P
Introduction to Machine Learning—Examples of Supervised Learning	Mohorianu	Thu	04 Mar 2021	16:00-17:00	P
Introduction to Machine Learning—Applying ML on RNAseq data	Mohorianu	Mon	08 Mar 2021	15:00-17:00	P

Machine learning gives computers the ability to learn without being explicitly programmed. It encompasses a broad range of approaches to data analysis with applicability across the biological sciences. In this section we will overview standard un-supervised and supervised approaches, exemplified first on theoretical, benchmark examples, and applied, in the last practical, on real RNA-seq data.

Block B

Bioinformatics Approaches to Omics Analysis

01	02	03	04
NGS	Sequence Alignment	GWASs	RNA-Seq

The *Bioinformatics Approaches to Omics Analysis, will introduce omics data and bioinformatics workflows used to process omic datasets with hands-on practice on genomic and transcriptomic data. We will go through the different stages of the omic data workflow, starting from the raw data, quality control, alignment, variant calling and analysis. Finally, genome-wide association studies are introduced, a popular approach in population studies which allow the identification of associations between single-nucleotide polymorphism (SNPs) loci and traits.*

01 Next Generation Sequencing (NGS)

Title	Lecturer	Day	Date	Time	Format
NGS data analysis: library preparation and quality control	Steif	Thu	21 Jan 2021	16:00-17:00	L
NGS data analysis: alignment and variant calling	Steif	Tue	26 Jan 2021	16:00-17:00	L
NGS practical: quality control, alignment and variant calling	Steif	Mon	01 Feb 2021	15:00-17:00	P

Lectures will provide an introduction to Next Generation Sequencing (NGS) technology, including library preparation and sequencing by synthesis. We will examine key file formats, learn how to assess sequencing data quality, and understand how reads aligned to a reference genome can be used to infer different types of genomic variants. The practical will provide a hands-on opportunity to perform quality control, alignment and variant calling.

02 Sequence Alignment

Title	Lecturer	Day	Date	Time	Format
Sequence Alignment	Brown	Thu	28 Jan 2021	16:00-17:00	L
Sequence Alignment	Brown	Tue	02 Feb 2021	16:00-17:00	P

This lecture and practical will cover the basic principles of sequence alignment, including similarity searching strategies such as BLAST and local and global multiple sequence alignments. The lecture will discuss different sequence alignment strategies and how they work, how sequence alignment is carried out in practice and some downstream applications. In the practical students will generate their own alignments and explore some of the potential applications.

03 Genome-Wide Association Studies (GWASs)

Title	Lecturer	Day	Date	Time	Format
Linking SNPs to disease and disease to biology	Day	Thu	11 Feb 2021	16:00-17:00	L
Using PLINK and R to find genetic associations	Day	Mon	22 Feb 2021	15:00-17:00	P

The lecture and practical will introduce the use of PLINK for generating genome-wide association estimates and how to handle imputed human genetic data. It will also cover the use of R for this type of study. The lecture will also cover studies design and a range of possible downstream uses of this sort of data. With an emphasis on how it can be applied to help us understand human health and disease.

04 RNA-Seq

Title	Lecturer	Day	Date	Time	Format
RNA-Seq analysis—General overview of RNAseq	Mohorianu	Tue	16 Feb 2021	16:00-17:00	L
RNA-Seq analysis—mRNAseq	Mohorianu	Thr	18 Feb 2021	16:00-17:00	L
RNA-Seq analysis—sRNAseq and integrating coding and non-coding RNAs	Mohorianu	Mon	01 Mar 2021	15:00-17:00	P

RNA sequencing (RNA-Seq) uses the NGS technology to identify and quantify RNA in a sample, thus allowing us to analyse the transcriptome. In this section we will overview the characteristics of RNAseq experiments and discuss their impact on computational analysis. The pipeline for identifying differentially expressed protein coding and non-protein coding genes will be presented in detail.

Block C

Pathway and Network Analysis of Biological Data

01	02
Gene-Set Enrichment Analysis	**Biological Networks**

In the last block of the course, *Pathway and Network Analysis of Biological Data, we will introduce ontologies and gene set enrichment analysis to link results obtained from the previous analyses back to biology and identify classes of genes or proteins that are over-represented in our results which may have an association with disease phenotypes. It will also introduce basic concepts of biological networks and their analysis with hands-on practice using Cytoscape, a widely used platform for Network Analysis.*

01 Gene Set Enrichment Analysis

Title	Lecturer	Day	Date	Time	Format
Gene-Set Enrichment Analysis	Mohorianu	Tue	09 Mar 2021	16:00-17:00	L

Gene set enrichment analysis is a method that identifies groups of genes or proteins that are over-represented in the analysis results. In this section we will focus on the interpretation of subsets of genes (obtained for example, from a differential expression analysis) using information from public databases - GO (term related to cellular components, molecular functions and biological processes), KEGG and Reactome pathways, and regulatory terms. This allows us to link our results back to biology.

02 Biological Networks

Title	Lecturer	Day	Date	Time	Format
Construction and Analysis of Biological Networks	Yadav	Thr	11 Mar 2021	16:00-17:00	L
Construction and Analysis of Biological Networks	Yadav	Mon	15 Mar 2021	15:00—17:00	P

This lecture will cover basic concepts and theory of Biological Networks, their types, and applications. Students will learn how to construct and visualise simple text and tabular data as a network, apart from understanding major network layout algorithms, visual styles and tips for effective visualisation. This will be followed by taking inputs from large high throughput datasets such as from protein-protein interaction and gene co-expression studies. The practical will focus on a hands-on experience in the use of Cytoscape, one of the most widely used open platforms for Network Analysis. Students will perform network construction and visualisation using their own datasets, followed by use of selected network analysis apps for clustering and detection of network motifs. Links to publicly available resources and hands-on exercises will be shared for further reading and practice.

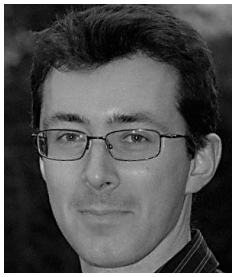
Lecturer Biographies

Dr Katy Brown

I am currently a postdoctoral research associate in the Firth Lab, part of the Virology Division in the Department of Pathology. I am a bioinformatician and my research involves identifying and characterising RNA viruses in RNA-seq datasets from bees, ants, wasps and mites. I previously worked at the University of Oxford as a computational genomics fellow and I completed my PhD on primate endogenous retroviruses at the University of Nottingham in 2014.

Dr Alexia Cardona

Dr Cardona leads training development of the University of Cambridge's Bioinformatics Training Programme. Her role involves the management of the different aspects of training including design, development, coordination and teaching of undergraduate and postgraduate training in Bioinformatics and Data Science. She is a leader in the ELIXIR international community, where together with the other leaders and partners drives the establishment of high-quality training in Data Science and Data Management for the Life Sciences. Dr Cardona is an advocate of participation in Communities of Practice and of women in leading and computational sectors which are currently underrepresented.



Dr Matt Castle

Matt Castle established the Graduate School of Life Sciences Biostatistics Initiative at the University of Cambridge in 2018 and currently leads the development and implementation of its training programmes and services. He has a PhD in mathematical epidemiology and spent a number of years as a post-doctoral researcher working with governments and NGOs on disease control strategies during his time as a senior modeller for the Epidemiology & Modelling Group. He has wide experience teaching on and developing material for a range of undergraduate and graduate courses in statistics, programming and mathematical biology at the University and for the Wellcome Trust. Matt also works with the Cambridge Centre for Teaching and Learning to support early career academics and researchers develop their pedagogical understanding and teaching skills as part of the Teaching Associates' Programme (TAP).

Dr Felix Day

Felix has studied for a Masters in Epidemiology from LSHTM and from MRC Epidemiology Unit, Cambridge. He has worked on large scale genetics and Mendelian Randomisation studies focused initially on obesity and related metabolic disorders. His recent work has focused on puberty timing and other reproductive phenotypes. This has included linking DNA damage repair to the timing of menopause. And the role of AMH on Polycystic Ovary Syndrome. These studies have also used reproductive outcomes as model traits for the identification of wider genetic effect; such as the impact of parent-of-origin-effect on the timing of menarche and the identification of a locus FADS1 which appears to be under selection in both modern and ancient populations.

Prof Richard Durbin

Richard Durbin is the Al Kindi Professor in the Department of Genetics, where he has been since 2017. Prior to that he was at the Wellcome Sanger Institute, where he had been since its founding in 1992. He has been involved in the bioinformatics and data analysis of many large genome sequencing projects, and he co-led the 1000 Genomes Project. He has broad research interests in computational genomics for genome sequence assembly, population sequence variation and genome evolution. Currently he is a partner in the Darwin Tree of Life project to sequence all species in Britain and Ireland, and studies speciation in Lake Malawi cichlid fish. He has also made many contributions to biological sequence analysis, including developing methods for sequence alignment using Hidden Markov models and suffix arrays, and genomic databases including Pfam, Ensembl, and TreeFam. Richard is a Foreign Member of the American Academy of Arts and Sciences, a Member of EMBO and a Fellow of the Royal Society.

Dr Sergio Martinez Cuesta

Sergio is starting a research group in bioinformatics at the interface between industry and academia exploring fundamental mechanisms of DNA epigenetics and protein degradation. He builds from his early training in experimental chemistry and biochemistry and the more recent bioinformatics PhD and postdoc positions to develop projects in collaboration with experimental colleagues sharing time between the AstraZeneca headquarters, the CRUK-CI in the biomedical campus and the Department of Chemistry in Cambridge. Sergio contributes as an editor of the emerging journal Frontiers in Bioinformatics and supports the growth of the European life science data infrastructure ELIXIR. Beyond research, he has a passion for team sports, languages and gastronomy.

Dr Irina Mohorianu

My background in Computer Science was streamlined towards Applied Data Science (Bioinformatics). During my PhD I studied small non coding RNAs (sRNAs), in plants and animals. During my postdoc I started to integrate mRNA and sRNA expression and became interested in Gene Regulatory Networks. In parallel I also continued working on developing new methods for the UEA sRNA Workbench. During my PhD, post-doc and as Head of Bioinformatics/ Scientific Computing at Cambridge Stem Cell Institute I continuously adapted statistical, Machine Learning and Data mining approaches to answer biological questions arising from experiments. My group is interested in both data analyses and method development applied to bulk and single cell multi-omics datasets. Integrative analyses are also combined with imaging analyses intertwined with other high throughput measurements.

Dr Adi Steif

Dr Steif is a Junior Research Fellow at Trinity College and postdoctoral fellow at the Cancer Research UK Cambridge Institute. Her research focuses on developing computational methods to analyse large genomic datasets in order to understand how cancers arise and evolve at the single cell level.

Dr Gita Yadav

Gita is a Lecturer at the Dept of Plant Sciences, University of Cambridge. She is also an R Instructor at the Cambridge Bioinformatics Training Facility. Gita's research is in the area of structural bioinformatics and complex networks, with applications in food security and conservation. She has a diverse educational background with a Ph.D. in Immunology, a Master's degree in Biomedical Research and a Graduate degree in Botany, from the University of Delhi, India. Gita has received several awards for her research and training efforts, including the Hamied Fellowship from the University of Cambridge, Exceptional Talent Award from the Royal Society of London, INSA Medal from the India National Science Academy, and the Women's Excellence Award from SERB, Govt of India.

Contact us:
Cambridge Bioinformatics Training
Craik-Marshall Building
Downing Site
University of Cambridge
Cambridge
CB2 3EB
United Kingdom

Email: grad.bioinfo@lifesci.cam.ac.uk
Telephone: +44 (0)1223 333614
Website: https://bioinfotraining.bio.cam.ac.uk/
Mailing list: https://lists.cam.ac.uk/mailman/listinfo/ucam-bioinfo-training
Follow us on Twitter: @BioInfoCambs

UNIVERSITY OF
CAMBRIDGE