AMSI AUSTRALIAN MATHEMATICAL SCIENCES INSTITUTE

Australian Government
Department of Education

THE UNIVERSITY OF SYDNEY

DAY 1 – Monday 2 December

STREAM A
14:15 – 15:00
15:20 – 16:30

Biology in the metaverse – a Virtual Reality tour

Presenters:
Philip Poronnik (University of Sydney)
Jim Cook (ICT Techlab University of Sydney)
Peter Thorn (University of Sydney)

Key words: basic biology, proteins, cells, virtual reality

Description: This workshop is designed to introduce you to the world of biology using immersive VR environments where you can explore biology fundamentals at different scales. You will learn about some interesting techniques in biology and how we generate large amounts of data that requires visualization and interpretation. We will also introduce you to basic VR workflows and how you can easily create your own worlds to tell data stories. The workshop will give you the chance to explore VR worlds, learn some interesting aspects of biology and consider how you might use VR in your future projects.

Requirements: None - we will provide VR headsets and appropriate reading materials etc.

Relevance: This is relevant to anyone with an interest in "all things science" who wants to appreciate the wonder and complexity of modern human biology.

DAY 1 – Monday 2 December

STREAM B
14:15 – 15:00
15:20 – 16:30

Enter the tidyverse with R and RStudio (Part A)

Presenters:
Kevin Wang, University of Sydney
Dr Garth Tarr, University of Sydney

Key words: statistical computing; R; tidyverse; data manipulation; data visualisation

Description: This workshop will familiarise you with the basics of R through the RStudio interface and the tidyverse suite of R packages. You will be introduced to modern approaches to data analysis and visualisation. The focus is on mastering basic skills and showing you where to go for help so you can undertake future analyses independently. By the end of this workshop you will know how to create and organise new "projects" in RStudio; read in data files; visualise data using the popular ggplot2 package; perform various data manipulation, summarisation and modelling tasks; and create reproducible reports for bioinformatics analysis pipelines.

In Part A of this workshop, we will first familiarise ourselves of the basics of R, e.g. loading in an Excel dataset, recognising variable types. We will be using the R Markdown documentation system, which allows us to execute codes, visualise output and writing a report. Time permitting, we will also start to learn the basics of data manipulations such as filtering of observations and selection of columns.

Some of the packages to be covered: rmarkdown, readr, readxl, voom, janitor and dplyr.

Requirements: You will need to bring your own laptop. Please make sure it has the latest version of R installed and the latest version of RStudio Desktop. Participants do not need to have existing knowledge of either R or RStudio.

Relevance: This workshop is relevant to anyone who is interested in learning more about R and how it can help streamline your data processing and analysis workflow. For example, if you currently spend a lot of time doing repetitive manual data manipulation tasks in Excel, you will benefit greatly from learning more about a statistical computing language such as R and the process of generating code for reproducible analyses. This workshop is also for people who might have learnt R a few years ago and is interested in upskilling in the recent advances, such as the RStudio interface and the tidyverse suite of packages (ggplot2, dplyr, readr, etc).

DAY 1 – Monday 2 December

STREAM C
15:20 – 16:30

Introduction to Unix and RNAseq processing

Presenters:
Dr Kitty Lo, University of Sydney
Dr Dario Strbenac, University of Sydney

Key words: Unix; computing basics; RNAseq

Description: Most bioinformatics tools are designed to be run from the command line hence the ability to run simple command line programs is an essential bioinformatics skill. This workshop will familiarise you with the basics of the Unix command line interface. We will show you how to navigate the file structure, run simple programs with arguments and open files. To keep it relevant to bioinformatics, we will demonstrate the samtools program and learn how to peer inside some common bioinformatic file formars (e.g. BAM file and fastq files)

Requirements: You will need to bring your own laptop.

Relevance: This workshop is relevant to students without any experience of the Unix command line who would like to gain a basic understanding of the Unix environment.

DAY 2 – Tuesday 3 December

STREAM A
13:30 – 15:00
15:30 – 17:00

Open Data Resources for Human Genomics Research

Presenters:
Associate Professor Jason Wong, School of Biomedical Sciences, University of Hong Kong
Dr Rebecca Poulos, Children's Medical Research Institute, University of Sydney

Key words: Genomics; Cancer; Databasese; Bioinformatics.

Description: Over the past decade, human genomics research has been driven by the generation of enormous quantities of data. There has been a concerted effort by the scientific community to make much of this data publicly available for unrestricted use in scientific research. A wide range of databases and web services have been developed to make use of this data more easily accessible. In this workshop, we will introduce some of the most popular publicly available databases and resources used by the genomics research community. The objective of this workshop is to enabling attendees to become familiar with how these resources can be accessed and how they can potentially be used for research. The first part of the workshop will be focused on general human genomics data resources such as UCSC genome browser, gnomAD, GTEx and ENCODE. The second part of the workshop will be specifically focused on cancer genomics data resources including the TCGA, Genomics Data Commons and cbioPortal.

Requirements: Participants will gain most benefit from this workshop if they have access to a laptop with a Wifi connection. There will be minimal assumed knowledge.

Relevance: This workshop will be relevant to those who are interested in doing genomics research, and who want to know how to access the many publicly-available datasets and databases online. The second part of this workshop will be specifically relevant to those working in cancer research.

DAY 2 – Tuesday 3 December

STREAM B
13:30 – 15:00
15:30 – 17:00

Enter the tidyverse with R and RStudio (Part B)

Presenters:
Kevin Wang, University of Sydney
Dr Garth Tarr, University of Sydney

Key words: statistical computing; R; tidyverse; data manipulation; data visualisation

Description: This workshop will familiarise you with the basics of R through the RStudio interface and the tidyverse suite of R packages. You will be introduced to modern approaches to data analysis and visualisation. The focus is on mastering basic skills and showing you where to go for help so you can undertake future analyses independently. By the end of this workshop you will know how to create and organise new “projects” in RStudio; read in data files; visualise data using the popular ggplot2 package; perform various data manipulation, summarisation and modelling tasks; and create reproducible reports for bioinformatics analysis pipelines.

In Part B of this workshop, we will focus on data cleaning and data visualisation. This type of tasks is where the tidyverse framework becomes one of the most powerful tools in data science. We will learn how to summarise data, converting between "wide" and "tall" data frames and also how to integrate different datasets. Using the techniques we learnt, we will massage the data into a suitable format and perform some statistical modelling. We will also introduce some powerful wrapper functions that can help us to write better and cleaner codes.

Some of the packages to be covered: tibble, broom, purrr, dplyr, tidyr and ggplot2.

DAY 2 – Tuesday 3 December

STREAM C
13:30 – 15:00
15:30 – 17:00

3D Genomics and Long-range Gene Regulations

Presenter:
Ye Zheng, Fred Hutchinson Cancer Research Center

Key words: three-dimensional chromatin organization, long-range gene regulation, statistical genomics analysis, computational tools.

Description: Chromatin is dynamically organized within the three-dimensional nuclear space in a way that allows efficient genome packaging while ensuring proper expression and replication of the genetic materials. In this workshop, we will go through the state-of-the-art 3D genomics technologies and focus on the role of statistical methods and computational tools in analyzing 3D genomics data. We will focus on introducing the standard processing pipeline as well as the widely used and fancy software in the field. Participants will have hands-on practical strategies to process 3D genomics data. Successful running of the complete pipeline and all software is not strictly required; instead, we will concentrate on the inference and interpretation of the results.

Requirements: You will need to bring your laptop and have the latest R and Python installed. Participants should be comfortable about running commands in terminal and have basic knowledge of Statistics. Participants are not expected to have any knowledge of 3D genomics.

Relevance: This workshop is relevant to anyone interested in learning three-dimensional chromatin structure, both from biotechnological and quantitative perspectives. The target audience can be anyone who came across 3C assays such as 3C, 4C, 5C, Hi-C, ChIA-PET, and HiChIP in literature and wants to learn more about them systematically. Or if you are simply curious about the three-dimensional chromatin structure and want to see some advanced experimental technologies and fancy quantitative analysis tools, this workshop is right for you!

DAY 3 – Wednesday 4 December

STREAM A
13:30 – 15:00
15:30 – 17:00

Single Cell RNA-seq Analysis

Presenters:
Hani Kim, University of Sydney
Yingxin Lin, University of Sydney
Dr Shila Ghazanfar, Cancer Research UK Cambridge Institute

Key words: Single-cell RNA-seq, data analysis; data integration

Description: Single-cell RNA-seq (scRNA-seq) is now widely used in many areas of biomedical research. Nonetheless, the analysis of scRNA-seq is often challenging and getting started can be a daunting task for beginners. This practical workshop is relevant to anyone who is interested in learning more about commonly used tools for scRNA-seq analysis in the R – Bioconductor environment. We will run through the process of analysing a scRNA-seq data collection from mouse fetal liver development from start to finish using open-source programs, including quality control, data integration, clustering analysis, differential expression analysis, pseudotime trajectory analysis and other popular single-cell downstream analysis.

Requirements: Participants are required to bring their own laptop. Basic R knowledge is encouraged but no previous single cell analytic experience is required.

Relevance: This is relevant to anyone who are interested in single-cell data analysis and want to learn commonly used tools for scRNA-seq analysis in the R – Bioconductor environment.

DAY 3 – Wednesday 4 December

STREAM B
13:30 – 15:00

Single cell RNA-seq analysis on the cloud

Presenters:
Associate Professor Joshua Ho, Hong Kong University
Andrian Yang, European Bioinformatics Institute
Xiunan Fang, Hong Kong University
Gordon Qian, Hong Kong University

Keywords: Spark; RNA-seq; big data

Description: Computational processing of large single cell RNA-seq data has many challenges, including the scalable processing of tens to hundreds of gigabytes of data, using memory and CPU intensive computational programs. This can be especially challenging if local computational resources are limited. *Falco* is a software bundle that enables bioinformatic analysis of large-scale transcriptomic data by utilising public cloud infrastructure. The framework currently provides supports for single cell RNA feature quantification, alignment and transcript assembly analyses. This workshop is a hands-on practical session on using Falco to run scalable bioinformatics analysis of single cell RNA-seq data.

Requirements: You will need to have an Amazon Web Service (AWS) account. Experience with working in the Unix command line environment is necessary.

Relevance: This workshop is relevant to anyone who are keen to explore the use of cloud computing for bioinformatics analysis, especially for single cell RNA-seq analysis.

DAY 3 – Wednesday 4 December

STREAM B
15:30 – 17:00

Pitfalls and roadblocks in single-cell analyses

Presenters:
Dr John Marioni, EMBL-EBI, University of Cambridge
Dr Shila Ghazanfar, Cancer Research UK Cambridge Institute

Key words: statistics; transcriptomics; normalisation; data integration; mean-variance effects

Description: In this workshop, we will focus on the bleeding edge of single-cell genomics, discussing some of the pitfalls and roadblocks that afflict many analyses. We will begin by highlighting some of the analysis steps that we find the most challenging and time-consuming and outline some things to be aware of that might indicate good or poor performance. Attendees will be encouraged to consider what analytical challenges they face in single-cell analyses and, ideally, to share with the group how they typically overcome these challenges.

Requirements: Good knowledge and experience of analysing large-scale and complex genomics datasets.

Relevance: This workshop is relevant to those who want additional hints and tips about the analyses of large and complex genomics datasets. It is ideally suited for those familiar with R / Bioconductor and state-of-the-art analyses approaches.

DAY 3 – Wednesday 4 December

STREAM C
13:30 – 15:00
15:30 – 17:00

Gene-expression analysis with RNA sequence data using R

Presenters:
Dr Atefeh Taherian Fard, Australian Institute for Bioengineering and Nanotechnology, University Queensland
Ms. Huiwen Zheng, Australian Institute for Bioengineering and Nanotechnology, University Queensland
Associate Professor Jessica Mar, Australian Institute for Bioengineering and Nanotechnology, University Queensland

Key words: R, RStudio, RNA-seq, differential expression; data visualisation, DESeq2 and pathway analysis

Description: In this workshop, you will learn how to analyse and explore RNA-seq count data. This hands-on workshop will cover basic steps in gene expression data analysis, including quality assessment, normalisation, differential gene expression testing, pathway over-representation analysis and visualisation. By the end of this workshop, you will be able to utilise the analysis workflow for your own RNA-seq data.

Requirements: Participants must bring their own laptop and make sure that it has the latest version of R and RStudio installed. Experience in using R and RStudio is desired but not required.

Relevance: This workshop is relevant to anyone who is interested in learning how to present RNA-seq data in an informative and engaging way, or applying different statistical methods, to understand the data and interpret the result using R.

DAY 4 – Thursday 5 December

STREAM A
13:30 – 15:00
15:30 – 17:00

Imputation and data quality control for proteomics data

Presenter:
Professor Pei Wang, Icahn School of Medicine at Mount Sinai, USA

Key words: proteomics, imputation

Description: Due to the dynamic nature of the mass spectrometry (MS) instruments, analyzing MS based proteomics data requires customized tools for routine preprocessing such as normalization, outlier detection/filtering, and batch correction. Moreover, proteomics data often contains substantial missing values. These together impose great challenges to data analyses. Specifically, many tools and methods, especially those for high dimensional data, often cannot deal with missing values directly. Furthermore, missing in proteomics data are not missing-at-random. Thus simply ignoring missing values or imputing them with constants will lead to biased results. In this talk, I will share a suite of preprocessing and imputation methods/tools for handling proteomics data. A specific focus will be given to an imputation method, DreamAI, which was resulted from an NCI-CPTAC Proteomics Dream Challenge that was carried out to develop effective imputation algorithms for proteomics data through crowd learning. DreamAI, is based on ensemble of six different imputation methods. The favorable performance of DreamAI over existing tools was demonstrated on both simulated and real data sets. Follow-up analysis based on the imputed data by DreamAI revealed new biological insights, suggesting this new tool could enhance the current data analysis capabilities in proteomics research.

Requirements: You will need to bring your own laptop. Please make sure it has the latest version of R installed.

Relevance: This workshop is relevant to anyone who is interested in analysing data from mass spectrometry based proteomics experiment.

DAY 4 – Thursday 5 December

STREAM B
13:30 – 15:00

Computational analysis for biological discovery from (phospho)proteomic data

Presenter:
Dr Pengyi Yang, University of Sydney

Keywords: Proteomics, data mining

Description: Mass spectrometry (MS) has become a well-established technology for global profiling of proteome and phosphoproteome in cells and tissues. Sophisticated computational methods are required for making sense of the data generated from MS-based proteomics and phosphoproteomics. In relation to professor Pei Wang's talk/workshop, which covers the methods/tools for preprocessing of MS-based proteomic data, this talk/workshop will introduce computational methods/tools for identifying kinases, substrates, and signalling and gene pathways that are regulated in different experimental conditions, assays, and time-series from large-scale proteomic and phosphoproteomic data. Hands-on demonstrations will be given in the workshop in which various computational methods/tools will be introduced through example applications to several proteomic and phosphoproteomic datasets using R programming environment.

Requirements: You will need to bring your own laptop. Please make sure it has the latest version of R installed.

Relevance: This workshop will introduce advanced computational methods for anyone interested in understanding the phosphoproteome from MS data.

DAY 4 – Thursday 5 December

STREAM B
15:30 – 17:00

Introduction to Proteomics

Presenters:
Ben Crossett, University of Sydney
David Maltby, University of Sydney
Angela Connolly, University of Sydney

Key words: mass spectrometry, proteomics

Description:
15:30 – 15:50
Proteomics 101: how to identify a protein by mass spectrometer

15:50 – 16:10
Half the class have a tour (hopefully inside) of SydneyMS, while half the class have a stab at PMF in demo laptops set up in the room.

16:10 – 16:30
As above but groups switch over.

Requirements: No equipment or knowledge required

Relevance: Students who have little knowledge of mass spectrometry will benefit from this intro workshop. This workshop also presents an opportunity for students to tour a workshop MS facility.

DAY 4 – Thursday 5 December

STREAM C (BioC)
13:30 – 15:00

R and Bioconductor for Genomic Analysis

Presenter:
Professor Martin Morgan, Roswell Park Comprehensive Cancer Center, Buffalo, New York, USA

Key words: Bioinformatics; R; gene expression; annotation; data management.

Description: This workshop will introduce you to the Bioconductor collection of R packages for statistical analysis and comprehension of high-throughput genomic data. The emphasis is on data exploration, using RNA-sequence gene expression experiments as a motivating example. How can I access common sequence data formats from R? How can I use information about gene models or gene annotations in my analysis? How do the properties of my data influence the statistical analyses I should perform? What common workflows can I perform with R and Bioconductor? How do I deal with very large data sets in R? These are the sorts of questions that will be tackled in this workshop.

Requirements: You will need to bring your own laptop. The workshop will use cloud-based resources, so your laptop will need a web browser and WiFi capabilities. Participants should have used R and RStudio for tasks such as those covered in introductory workshops earlier in the week. Some knowledge of the biology of gene expression and of concepts learned in a first course in statistics will be helpful.

Relevance: This workshop is relevant to anyone eager to explore genomic data in R. The workshop will help connect 'core' R concepts for working with data (e.g., data management via data.frame(), statistical modelling with lm() or t.test(), visualization usiing plot() or ggplot()) to the special challenges of working with large genomic data sets. It will be especially helpful to those who have or will have their own genomic data, and are interested in more fully understanding how to work with it in R.

DAY 4 – Thursday 5 December

STREAM C (BioC)
15:30 – 17:00

Building a Bioconductor package

Presenters:
Dr Peter Hickey, Walter and Eliza Hall Institute for Medical Research
Dr Saskia Freytag, Harry Perkins Institute of Medical Research

Key words: R package; Bioconductor package; Dissemination; Open Source

Description: This workshop will answer the following questions:

- What is an R package?
- Why make an R package?
- How to make an R package?
- How can I share my R package?

Participants will create their first small R package and share it with the world through GitHub. Throughout this process we will point out important aspects of package development, such as documentation, testing and design principles. To conclude, we will discuss the Bioconductor submission process and some helpful tips and tricks to get through it painlessly.

Requirements: You will need to bring your laptop. For the workshop you will be assigned an AWS instance with a working RStudio version, that can be accessed via any up-to-date browser (Chrome, Firefox). Please also sign up to GitHub (https://github.com/), if you have not already got an existing account.

Relevance: Sharing bioinformatics development through software is the most effective way to increase the significance and reach of your work. The Bioconductor repository is a recognized platform to share R software relating to biological data. However, the creation of an R package and its sharing can be daunting for a first-time user. We will alleviate your fears and show you that your first package is within your reach.

DAY 5 – Friday 6 December

STREAM A (BioC)
13:30 – 15:30

Differential discovery in high-dimensional cytometry data

Presenter:
Helena L. Crowell, University of Zurich, Switzerland

Key words: mass cytometry; CyTOF; visualization; clustering; dimension reduction; differential analysis

Description: In this workshop, we will cover an R-based pipeline for differential analysis of (replicated, multi-condition) high-dimensional mass cytometry data, which is largely based on Bioconductor infrastructure, and includes: i) identification of cell subpopulations using a sequence of high-resolution clustering, consensus clustering, manual merging and annotation; and, ii) differential abundance (DA) and state (DS) analyses, in order to identify association of population abundances with a phenotype, or changes in signalling within populations. Alongside formal statistical analyses, we will perform exploratory data analysis at each step, such as reporting on various clustering and differential testing results through dimensionality reduction, heatmaps of aggregated signal etc.
*The workshop will closely follow Nowicka et al.'s "CyTOF workflow: differential discovery in high-throughput high-dimensional cytometry datasets" (F1000Research, 2017), available here.

Requirements:
Technical: You will need to bring your own laptop. The workshop will use cloud-based resources, so your laptop will need a web browser and WiFi capabilities.
Knowledge/competencies: Participants are expected to have basic-intermediate knowledge of R and some familiarity with Bioconductor's SingleCellExperiment class.

Relevance: The workshop presented here will equip participants with the expertise for diverse exploratory and differential analyses of high-dimensional cytometry data with complex experimental design, i.e., multiple cell subpopulations, samples (e.g. patients), and conditions (e.g. treatments). Furthermore, a large proportion of the analyses presented here are transferable to scRNA-seq, and the workshop may thus be of interest also to anyone who is interested in analysing replicated multi-condition scRNA-seq data.

DAY 5 – Friday 6 December

STREAM B (BioC)
13:30 – 15:30

Fluent genomics: a plyranges and tximeta case-study

Presenter:
Stuart Lee, Monash University, Walter and Eliza Hall Institute for Medical Research

Key words:
genomics, RNA-seq, tidyverse, R programming, Bioconductor

Description:
In this workshop, we will give an overview of how to perform exploratory analyses of genomic data using the grammar of genomic data transformation defined in the plyranges package. In the first half of the workshop, we will introduce the GRanges data structure and provide an overview of the core verbs for arithmetic, restriction, and aggregation of GRanges. In the second half of the workshop, we will work through case study of integrating differential expression and differential chromatin accessibility results from an experiment of macrophage cell lines. We will learn how to use the tximeta package for automatically preparing data from an RNA-seq experiment with correct reference annotations.

Requirements: You will need to bring your own laptop. The workshop will use cloud-based resources, so your laptop will need a web browser and WiFi capabilities. A familiarity with the basics of R/tidyverse would be a plus but is not strictly necessary.

Relevance:
This workshop will be beneficial to new learners of R who would like to understand more about Bioconductor and learners who are already familiar with the tidyverse suite of packages and would like to apply those concepts to bioinformatics data analysis. It is also recommended for biological scientists who would like to start looking at data from their own experiments but are not sure how to begin.

DAY 5 – Friday 6 December

STREAM C (BioC)
13:30 – 15:30

Reproducible bioinformatics

Presenter:
Dave Tang

Key words: Docker; Conda; Bioconda; RStudio Server; Reproducibility; Project management

Description: This workshop will discuss guidelines for ensuring reproducibility in bioinformatic data analysis and demonstrate how we can adhere to these guidelines through the use of various computational tools. You will be introduced to Conda and Docker and shown how they can be used to simplify the deployment of bioinformatics tools and create isolated software environments ensuring that analyses can be reproduced. The workshop will also discuss approaches for organising computational projects using the workflowr R package. By the end of the workshop, you will have learned some ideas behind carrying out reproducible research and can better communicate and share your work in a reproducible manner.

Requirements: You will need to bring your own laptop. Please make sure it has the latest version of R and RStudio Desktop installed. In addition, please install the latest versions of Miniconda and Docker. Some command line experience will be helpful but not required. Further instructions available from
https://github.com/davetang/reproducible_bioinformatics

Relevance: One of the most important aspects of scientific research is that someone else can reproduce your work. Even if a complex bioinformatics analysis is thoroughly described in the supplementary material of a paper and all raw data is provided, this doesn't guarantee that other researchers can reproduce your work. This workshop is relevant to anyone who is interested in learning how to work in a manner that promotes reproducibility. In most cases, the person trying to reproduce your work is your future self. If you have looked back on your previous analyses and had trouble figuring out what you had done, this workshop is for you.