AMSI AUSTRALIAN MATHEMATICAL SCIENCES INSTITUTE

Australian Government
Department of Education

THE UNIVERSITY OF SYDNEY

THANK YOU TO THE AMSI BIOINFOSUMMER 2019 SPONSORS

JOIN THE CONVERSATION ON SOCIAL MEDIA

@DiscoverAMSI #BioInfoSummer

@DiscoverAMSI @bioinfosummer #BioInfoSummer

AMSI BioInfoSummer 2019

Charles Perkins Centre
The University of Sydney

Monday 2 – Friday 6 December

Local organising committee:

Jean Yang, The University of Sydney (Event Director)

Ellis Patrick, The University of Sydney (Event Director)

Kitty Lo, The University of Sydney

Lake-Ee Quek, The University of Sydney

Mengbo Li, The University of Sydney

Rebecca Poulos, Children's Medical Research Institute

Bobbie Cansdale, The University of Sydney

AMSI BioInfoSummer Standing Committee:

Matt Ritchie, Walter and Eliza Hall Institute of Medical Research (Chair)

Nicola Armstrong, Murdoch University

Tim Brown, Australian Mathematical Sciences Institute

Mike Charleston, University of Tasmania

Gary Glonek, The University of Adelaide

Ville-Petteri Makinen, The University of Adelaide

Jessica Mar, The University of Queensland

Alicia Oshlack, Murdoch Children's Research Institute

Tony Papenfuss, Walter and Eliza Hall Institute of Medical Research

Ellis Patrick, The University of Sydney

Chloe Pearse, Australian Mathematical Sciences Institute

David Powell, Monash University

Mat Simpson, Queensland University of Technology

Jean Yang, The University of Sydney

Australian Mathematical Sciences Institute:

Chloe Pearse, Research and Higher Education Program Manager

Angela Coughlin, Research and Higher Education Project Coordinator

Michael Shaw, Multimedia Manager

Francesca Hoban Ryan, Project Administration Assistant

FREE PUBLIC LECTURE AS PART OF AMSIBIOINFOSUMMER19

THE BRIGHT FUTURE OF APPLIED STATISTICS

PROFESSOR RAFAEL IRIZARRY
HARVARD UNIVERSITY

6PM – 7.30PM THURS 5 DEC

CPC AUDITORIUM THE UNIVERSITY OF SYDNEY

REGISTER BIS.AMSI.ORG.AU/PUBLIC-LECTURE

DAY 1 – Monday 2 December

INTRODUCTION TO BIOINFORMATICS

10:00-11:00	**REGISTRATION**
11:00-11:30	Conference Opening
11:30-12:30	Bioinformatic the discipline, versus the career Professor Melanie Bahlo, The Walter and Eliza Hall Institute of Medical Research
12:30-14:15	**WELCOME LUNCH / CHOOSE MATHS LUNCH EVENT (catered)**
14:15-15:00	**WORKSHOPS**
Stream A	Biology in the metaverse – a Virtual Reality tour Professor Philip Poronnik, Jim Cook, Professor Peter Thorn, The University of Sydney
Stream B	An introduction to statistics in the omics era Associate Professor Gary Glonek, The University of Adelaide
15:00-15:20	**AFTERNOON TEA (catered)**
15:20-16:30	**WORKSHOPS**
Stream A	Biology in the metaverse – a Virtual Reality tour Professor Philip Poronnik, Jim Cook, Professor Peter Thorn, The University of Sydney
Stream B	Enter the tidyverse with R and RStudio (Part A) Kevin Wang, Dr Garth Tarr, The University of Sydney
Stream C	Introduction to UNIX and RNA-seq processing Dr Kitty Lo, Dr Dario Strbenac, The University of Sydney
16:30-17:30	Saving the Tasmanian devil from extinction Professor Kathy Belov AO FRSN, The University of Sydney
17:30-18:30	**WELCOME RECEPTION (catered)**

OPENING LECTURE: BIOINFORMATIC THE DISCIPLINE, VERSUS THE CAREER

Professor Melanie Bahlo, The Walter and Eliza Hall Institute of Medical Research

There is no doubt that Bioinformatics/Computational Biology is now a vital component of modern biological research, whether this be in industry or academia, in biomedical or agricultural research or in the diagnostic setting. There is still an increasing need for people trained in these areas. This talk will examine some of the many guises of bioinformatics training and careers and try and identify some future trends. This will also involve some reflection of how things have changed over the last 20 years.

Professor Melanie Bahlo is a bioinformatician/statistical geneticist with over 20 years' experience working on the discovery of the genetic basis of human diseases, with a focus on neurological disorders. Leading the Statistical Genetics laboratory at The Walter and Eliza Hall Institute of Medical Research since 2007, her work combines the development of novel methods and successful applications in monogenic and complex diseases to identify and understand biological mechanisms perturbed in diseases. She co-established the WEHI Population Health and Immunity Division and has recently been promoted to the position of Healthy Development and Ageing Theme Leader.

She has won two Australian science awards including the AAS Moran Medal (2009) and the Genetics Society of Australasia's Ross Crozier Medal for mid-career researchers (2015).

WORKSHOP: BIOLOGY IN THE METAVERSE – A VIRTUAL REALITY TOUR

Professor Philip Poronnik, University of Sydney
Jim Cook, ICT Techlab University of Sydney
Professor Peter Thorn, University of Sydney

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.

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

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.

AN INTRODUCTION TO STATISTICS IN THE OMICS ERA

Associate Professor Gary Glonek, The University of Adelaide

Statistical concepts and methods play an important role in bioinformatics. The scale and complexity of typical bioinformatics data sets poses both challenges and opportunities for existing statistical theory. In this talk, basic statistical concepts will be reviewed and their use in bioinformatics applications will be illustrated.

Gary is a lecturer in statistics and the Head of the School of Mathematical Sciences at the University of Adelaide. His research interests are in statistics, especially with applications in bioinformatics. He is also interested in applied statistics and has undertaken consultancies across a wide range of areas, including road safety, wine quality and healthcare policy.

WORKSHOP: ENTER THE TIDYVERSE WITH R AND RSTUDIO (PART A)

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

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.

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

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).

WORKSHOP: INTRODUCTION TO UNIX AND RNASEQ PROCESSING

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

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)

Key words: Unix; computing basics; RNAseq

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.

SAVING THE TASMANIAN DEVIL FROM EXTINCTION

Professor Kathy Belov AO FRSN, The University of Sydney

Twenty years ago, a new disease emerged in Tasmania that threatened the iconic Tasmanian Devil with extinction. We determined that the disease was caused by a contagious cancer that was spread as an allograft by biting. The tumour spread quickly due to low levels of genetic diversity in the species and the tumours capacity to evade the immune system. Over 85% of the species was lost. Yet – although predicted, extinction has not occurred. Both devils and tumours evolved. The age structure of devil populations changed. Devils persisted in the wild, albeit in small, isolated populations. I will discuss the role that genomics has played in understanding devils and the disease. I will explain how we have used genomics to manage genetic diversity within Australia's largest captive breeding program and how we are now using insurance population animals for genetic rescue of populations in the wild. Beyond that, I will talk about how we are leveraging this approach to conserve an additional 50 threatened Australian species.

Professor Kathy Belov is a Professor of Comparative Genomics in the School of Life and Environmental Sciences in the Faculty of Science at the University of Sydney. Kathy's research expertise is in the area of comparative genomics and immunogenetics of Australian wildlife, including Tasmanian devils and koalas, two iconic species that are threatened by disease processes. Kathy's research team has participated in a range of marsupial and monotreme genome projects where they have gained insights into genes involved in immunity and defense, including platypus venom genes and novel antimicrobial peptides in the pouch. Kathy has published over 150 peer reviewed papers, including papers in Nature, Proceedings of the National Academy of Science and PLoS Biology. Kathy has received two Eureka awards, the Crozier medal from the Genetics Society of Australasia and the Fenner medal from the Australian Academy of Science for her research. She is currently the immediate past president of the Genetics Society of Australasia and a Fellow of the Royal Society of NSW.

Kathy is also the Pro-Vice-Chancellor (Global Engagement) at the University of Sydney. In this position she takes responsibility for managing the development and execution of the University's global engagement strategy. Kathy is passionate about mentoring others, particularly women in STEM.

DAY 2 – Tuesday 3 December

EPIGENETICS / GENOMICS

08:45-09:00	**REGISTRATION**
09:00-09:45	Epigenetics Demystified Associate Professor Clare Stirzaker, Garvan Institute of Medical Research
09:45-10:30	Accurate identification of mRNA alternative splicing using Oxford Nanopore sequencing Dr Heejung Shim, Melbourne Integrative Genomics (MIG), University of Melbourne
10:30-11:00	**MORNING TEA (catered)**
11:00-11:45	FreeHi-C: high fidelity Hi-C data simulation for benchmarking and data augmentation Ye Zheng, Fred Hutchinson Cancer Research Center
11:45-12:00	Diversity in STEM Chloe Pearse, Australian Mathematical Sciences Institute
12:00-13:30	**DIVERSITY IN STEM LUNCH EVENT (catered)**
13:30-15:00	WORKSHOPS ***Sponsored by Australian BioCommons***
Stream A	Open data resources for human genomics research Associate Professor Jason Wong, University of Hong Kong Dr Rebecca Poulos, Children's Medical Research Institute
Stream B	Enter the tidyverse with R and RStudio (Part B) Kevin Wang, Dr Garth Tarr, The University of Sydney
Stream C	3D Genomics and Long-range Gene Regulations Ye Zheng, Fred Hutchinson Cancer Research Center
15:00-15:30	**AFTERNOON TEA (catered)**
15:30-17:00	WORKSHOPS ***Sponsored by Australian BioCommons***
Stream A	Open data resources for human genomics research Associate Professor Jason Wong, University of Hong Kong Dr Rebecca Poulos, Children's Medical Research Institute
Stream B	Enter the tidyverse with R and RStudio (Part B) Kevin Wang, Dr Garth Tarr, The University of Sydney
Stream C	3D Genomics and Long-range Gene Regulations Ye Zheng, Fred Hutchinson Cancer Research Center
17:00-18:30	**COMBINE CAREERS EVENING (catered)**

EPIGENETICS DEMYSTIFIED

Associate Professor Clare Stirzaker, Garvan Institute of Medical Research

Epigenetics plays a critical role in normal cellular differentiation and development. Epigenetic regulation of the genome is governed by many facets of epigenetic regulation, which include DNA methylation, chromatin modifications, nucleosome positioning and higher order chromatin structure. It is well established that normal epigenetic processes are commonly disrupted in disease, including cancer, contributing to alterations in the transcriptome and deregulation of cellular pathways. Understanding the complex relationship between DNA methylation, chromatin modifications and underlying DNA sequence is a major focus in cancer biology. Genome-wide distribution of DNA methylation, post-translational histone modifications and chromatin structure are being extensively mapped by next-generation sequencing technologies, revealing the dynamic interplay between the epigenetic marks and how they are altered in disease states. Genetic changes to epigenetic modifiers can lead to aberrations of the normal epigenome, revealing the mechanisms underpinning altered cellular pathways in disease, and highlighting prospects for future epigenetic therapies.

Clare completed her Bachelor of Science majoring in Biochemistry and Microbiology, and her Bachelor of Science with First Class Honours in Molecular Biology, at the University of Cape Town, South Africa, before completing her PhD at Macquarie University, Sydney, graduating in 1990.

Clare became fascinated by the field of Epigenetics and joined the group of Prof Susan Clark as a post-doc at the Kanematsu Laboratories, Royal Prince Alfred Hospital, in Sydney and later, at the Sydney Cancer Centre at Sydney University. The group moved to the Garvan Institute of Medical Research in 2004, where an Epigenetics Group was established within the Cancer Program. Clare has since established her own group, which is interested in understanding "Epigenetic Deregulation in Cancer."

Clare has played a major role in delivery of many of the milestones in epigenetic research. She has made highly significant contributions to the field of DNA methylation and epigenetic deregulation in cancer and has also played an integral role in developing new epigenetic technologies that have underpinned many of the seminal findings of the group.

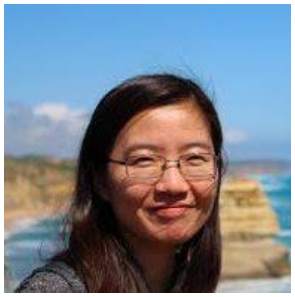
Clare is involved in a number of collaborative projects which include an NBCF project grant 'Enabling Clinical Epigenetic Diagnostics: The Next Generation of Personalised Breast Cancer Care' and an NHMRC project grant with Monash University 'Defining Epigenetic Changes in Prostate Cancer Stroma.'

ACCURATE IDENTIFICATION OF MRNA ALTERNATIVE SPLICING USING OXFORD NANOPORE SEQUENCING

Dr Heejung Shim, Melbourne Integrative Genomics (MIG), University of Melbourne

The pre-mRNA alternative splicing events enable a single gene to produce different proteins in eukaryotes, and they have been shown to affect various gene functions, and eventually disease. The alternative splicing can not only add or skip entire exons, but can also vary exon boundaries by

selecting different nucleotides as splice sites. Oxford Nanopore sequencing produces long reads that have natural advantages for characterising alternative splicing events. Nanopore sequencing records changes in electrical current when a DNA or RNA strand is traversing through a pore. This raw signal, known as a squiggle, is then basecalled by computational methods. However, due to the high error rate in the basecalling process, it is challenging to accurately identify exon boundaries using the basecalled sequences. In this talk, I will introduce new methods that use the squiggle to characterise the exon boundaries, in addition to the basecalled sequences.



Dr. Heejung Shim is a Group Leader in the Melbourne Integrative Genomics (MIG) and Lecturer in the School of Mathematics and Statistics at the University of Melbourne. She completed her BS in Mathematics (with a double major in Computer Science and Engineering) from the POSTECH, and her PhD in Statistics from the University of Wisconsin at Madison, advised by Prof. Bret Larget. She did a postdoc at the University of Chicago working with Prof. Matthew Stephens. Previous to her position at the University of Melbourne, she was an Assistant Professor in the Department of Statistics at the Purdue University for two years. Currently she retains an affiliation with Purdue as an Adjunct Assistant Professor.

FREEHI-C: HIGH FIDELITY HI-C DATA SIMULATION FOR BENCHMARKING AND DATA AUGMENTATION

Ye Zheng, Fred Hutchinson Cancer Research Center

Ability to simulate realistic high-throughput chromatin conformation (Hi-C) data is foundational for developing and benchmarking statistical and computational methods for Hi-C data analysis. We propose FreeHi-C, a data-driven Hi-C simulator for simulating and augmenting Hi-C datasets. FreeHi-C employs a non-parametric strategy for estimating interaction distribution of genome fragments from a given sample and simulates Hi-C reads from interacting fragments. Data from FreeHi-C exhibit higher fidelity to the biological Hi-C data compared with other tools in its class. FreeHi-C not only enables benchmarking a wide range of Hi-C analysis methods but also boosts the precision and power of differential chromatin interaction detection methods while preserving false discovery rate control through data augmentation.

I received my B.E in Statistics at Renmin University of China before starting my doctoral training at the University of Wisconsin – Madison in Fall 2014. I received my Ph.D. in Statistics with a doctoral minor in Quantitative Biology under the supervision of Professor Sündüz Keleş in August 2019. I am inherently drawn to problems that are at the interface of statistical, biological, and biomedical sciences. My research thus far concentrated on developing statistical and computational methods for studying three-dimensional chromatin organization and long-range regulatory interactions in DNA. I will join Professor Raphael Gottardo's group at Fred Hutchinson Cancer Research Center as a Postdoctoral Research Fellow to further investigate the genomic regulation mechanism from the single-cell perspective.

WORKSHOP: OPEN DATA RESOURCES FOR HUMAN GENOMICS RESEARCH

Associate Professor Jason Wong, School of Biomedical Sciences, University of Hong Kong

Dr Rebecca Poulos, Children's Medical Research Institute, University of Sydney

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.

Key words: Genomics; Cancer; Databasese; Bioinformatics.

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.

WORKSHOP: ENTER THE TIDYVERSE WITH R AND RSTUDIO (PART B)

Kevin Wang, University of Sydney

Dr Garth Tarr, University of Sydney

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.

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

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).

WORKSHOP: 3D GENOMICS AND LONG-RANGE GENE REGULATIONS

Ye Zheng, Fred Hutchinson Cancer Research Center

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.

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

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

SINGLE CELL OMICS

08:45-09:00	**REGISTRATION**
09:00-09:45	Rethinking the atlas paradigm: Moving from descriptive to predictive computational biology Professor Christine Wells, Centre for Stem Cell Systems
09:45-10:30	Understanding cell fate decisions using single cell genomics Dr John Marioni, EMBL-EBI, Cambridge University
10:30-11:00	**MORNING TEA (catered)**
11:00-11:30	Multidimensional single cell analysis of the tumour microenvironment Associate Professor Alex Swarbrick, Garvin Institute of Medical Research
11:30-11:50	Investigating higher order interactions in single cell data with scHOT Dr Shila Ghazanfar, Cancer Research UK Cambridge Institute
11:50-12:10	Why bulk samples (still) matter for gene expression analysis Associate Professor Jessica Mar, Australian Institute for Bioengineering and Nanotechnology
12:10-12:30	Scalable bioinformatics methods for single cell data Associate Professor Joshua Ho, University of Hong Kong
12:30-13:30	**LUNCH**
13:30-15:00	WORKSHOPS ***Sponsored by The Westmead Institute for Medical Research***
Stream A	Single Cell RNA-seq Analysis Hani Kim, Yingxin Lin, University of Sydney Dr Shila Ghazanfar, Cancer Research UK Cambridge Institute
Stream B	Single cell RNA-seq data analysis on the cloud Associate Professor Joshua Ho, University of Hong Kong
Stream C	Gene expression analysis with RNA-Seq data using R Associate Professor Jessica Mar, Dr Atefeh Taherian Fard, Huiwen Zheng, Australian Institute for Bioengineering and Nanotechnology
15:00-15:30	**AFTERNOON TEA (catered)**
15:30-17:00	WORKSHOPS ***Sponsored by The Westmead Institute for Medical Research***
Stream A	Single Cell RNA-seq Analysis Hani Kim, Yingxin Lin, University of Sydney Dr Shila Ghazanfar, Cancer Research UK Cambridge Institute
Stream B	Pitfalls and roadblocks in single-cell analyses Dr John Marioni, EMBL-EBI, Cambridge University Dr Shila Ghazanfar, Cancer Research UK Cambridge Institute
Stream C	Gene expression analysis with RNA-Seq data using R Associate Professor Jessica Mar, Dr Atefeh Taherian Fard, Huiwen Zheng, Australian Institute for Bioengineering and Nanotechnology

RETHINKING THE ATLAS PARADIGM: MOVING FROM DESCRIPTIVE TO PREDICTIVE COMPUTATIONAL BIOLOGY

Professor Christine Wells, Centre for Stem Cell Systems

The curation and description of biological systems underpins how we classify and organize knowledge about ourselves and our world, allows us to categorise new parts, and is the cornerstone to understanding relationships between molecules and phenotypes. As we develop the tools to isolate individual cells, and even compartments of those cells the scale of the catalogue increases. What are the principles that are common to atlas projects that allow insights to go beyond the catalogue? What are the common pitfalls that need to be relearned through each iteration of the molecular atlas? And what is ahead for the field – will the atlas reach information saturation, and what are the big questions left for computational biologists to consider?

Professor Christine Wells is the Director of the University of Melbourne Centre for Stem Cell Systems. She is the Deputy Program Lead for Stem Cells Australia, an Australian Research Council funded $24M special research initiative that brings together leading Australian stem cell scientists. She graduated from the University of QLD in 2004 (PhD), and over the following 15 years has worked with international consortia including FANTOM (RIKEN, Japan); Functional Glycomics (USA); Project Grandiose (Canada) and Leukomics (UK).

Christine is a systems biologist interested in tissue injury and repair. She leads a program of research in biological data integration and visualization, including method development leading to gene discovery and characterization of stem cell subsets and innate immune cells. Christine was the first to discover a role for the C-type lectin Mincle in host responses to infection, and has since characterised a role for Mincle in ischemic injury. She has published over 110 scientific journal articles, in the leading scientific literature, including landmark studies mapping gene architecture and function. She has developed several open source software programs, including the Stemformatics.org stem cell collaboration resource, which has a global audience, and which hosts a large compendium of curated stem cell data. This resource is used to generate definitive molecular signatures of stem cell subsets and their differentiated progeny, benchmark in vitro differentiated stem cell attributes, particularly of pluripotent-derived myeloid cells.

UNDERSTANDING CELL FATE DECISIONS USING SINGLE CELL GENOMICS

Dr John Marioni, EMBL-EBI, Cambridge University

John Marioni obtained his PhD in Applied Mathematics in the University of Cambridge in 2008 and did his postdoctoral research in the Department of Human Genetics, University of Chicago. He joined EMBL-EBI as Research Group Leader in Computational and Evolutionary Genomics in 2010. His group develops the computational and statistical tools necessary to exploit high-throughput genomics data, with the aim of understanding the regulation of gene expression and modelling developmental and evolutionary processes. Within this context, the Marioni group focuses on

understanding how the divergence of gene expression levels is regulated, using gene expression as a definition of the molecular fingerprint of individual cells to study the evolution of cell types, and modelling spatial variability in gene-expression levels within a tissue or organism. These three strands of research are brought together by single-cell sequencing technologies. John has a joint appointment at the Wellcome Trust Sanger Institute and the Cancer Research UK Cambridge Institute, which is part of the University of Cambridge.

MULTIDIMENSIONAL SINGLE CELL ANALYSIS OF THE TUMOUR MICROENVIRONMENT

Associate Professor Alex Swarbrick, Garvin Institute of Medical Research

Solid cancers are a complex 'ecosystem' of diverse cell types, whose heterotypic interactions play central roles in defining the aetiology of disease and its response to therapy. We used a multi-dimensional single cell genomics approach to characterise the tumour microenvironment in a unique cohort of early breast cancers.

Malignant cells showed remarkable intra-tumoural heterogeneity for canonical breast cancer features, such as intrinsic subtype, hormone receptor expression and transcriptional drivers. By integrating with the Nanostring DSP platform for spatial RNA profiling, we identify signatures to distinguish malignant cell clusters from benign and morphologically normal epithelial cells.

Cancer Associated Fibroblasts (CAFs) were found in at least two states: a myofibroblast-like subset and an inflammatory-mediator subset. Distinct transcription factor networks regulated these polarised states. We show distinct functions for these subsets.

We applied CITE-Seq to measure >150 cell surface immune markers and checkpoint proteins simultaneous to RNA-Sequencing. We resolve the tumour-immune milieu with high precision and generate new transcriptional signatures of breast tumour-infiltrating leukocytes.

To track lymphocyte clonal dynamics through space and time, we developed a novel method to permit simultaneous full-length lymphocyte receptor- and short-read RNA-sequencing at single cell resolution. We observe clonal expansion and trafficking of CD4+ and CD8+ T lymphocytes between the lymph nodes, blood and tumor of patients.

This data provides by far the most extensive insights into the cellular landscape of breast cancer and will reveal new biomarkers and opportunities for stromal- and immune-based therapy.

Alex graduated with a Bsc (Hons I) in Molecular and Cellular Biology from UNSW in 1995. After obtaining his PhD in 2003 he undertook postdoctoral training with Nobel laureate J. Michael Bishop at the University of California, San Francisco, supported by a CJ Martin Travelling Fellowship from the NHMRC.

In 2008 Alex established the Tumour Progression Laboratory in the Garvan Institute and in 2012 was appointed co-Head of the Breast Translational ncology Program in the newly commissioned Kinghorn Cancer Centre. Alex is an Associate Professor at UNSW and an NHMRC Senior Research Fellow.

His lab uses single cell 'omics analysis of clinical tissue cohorts and patient-avatar models to discover novel treatment strategies for solid cancers, including breast, prostate, melanoma and neuroblastoma.

Alex is the convenor of the Lorne Cancer Conference, Australia's pre-eminent multi-disciplinary cancer research conference & the Australian Translational Breast & Prostate Cancer Symposium. He serves on the Cancer Research Committee of the Cancer Council NSW.

INVESTIGATING HIGHER ORDER INTERACTIONS IN SINGLE CELL DATA WITH scHOT

Dr Shila Ghazanfar, Cancer Research UK Cambridge Institute

Single-cell RNA-sequencing has transformed our ability to examine cell fate choice. For example, in the context of development and differentiation, computational ordering of cells along 'pseudotime' enables the expression profiles of individual genes, including key transcription factors, to be examined at fine scale temporal resolution. However, while cell fate decisions are typically marked by profound changes in expression, many such changes are downstream of the initial cell fate decision. By contrast, more subtle changes in patterns of correlation and higher order interactions between genes across pseudotime have been associated with the fate choice itself.

We describe a novel approach, scHOT – single cell Higher Order Testing – which provides a flexible and statistically robust framework for identifying changes in higher order interactions among genes. scHOT is general and modular in nature, can be run in multiple data contexts such as along a continuous trajectory, between discrete groups, and over spatial orientations; as well as accommodate any higher order measurement such as variability or correlation. scHOT meaningfully adds to first order effect testing, such as differential expression, and provides a framework for interrogating higher order interactions from single cell data.

Dr. Shila Ghazanfar is a Royal Society Newton International Fellow and Research Associate working at the Cancer Research UK Cambridge Institute. She completed her PhD in statistical bioinformatics at The University of Sydney in the School of Mathematics and Statistics. Her current research interests are in the statistical analysis of data arising from high throughput sequencing technologies such as single cell RNA-Seq and spatially resolved single cell transcriptomics in various research contexts.

WHY BULK SAMPLES (STILL) MATTER FOR GENE EXPRESSION ANALYSIS

Associate Professor Jessica Mar, Australian Institute for Bioengineering and Nanotechnology

For beginner bioinformaticians, you may be inclined to believe that all RNA-sequencing now revolves around single cells. While advances in single-cell next-generation sequencing technologies have created new opportunities for computational biology, it is useful to recognize that the equivalent data sets generated from bulk samples still have their place for bioinformatics and for understanding

biology. This talk outlines some of the reasons why bioinformaticians still need to pay attention to bulk RNA-sequencing data sets and in particular, why beginners should learn how to analyze gene expression data at both bulk and single-cell resolution.

Associate Professor Jessica Mar is a Group Leader at the Australian Institute for Bioengineering and Nanotechnology at The University of Queensland in Brisbane. The Mar group focuses on understanding variability in the transcriptome and how this informs regulation of cell phenotypes. Jess received her PhD in Biostatistics from Harvard University in 2008. She was a postdoctoral fellow at the Dana-Farber Cancer Institute in Boston (2008-2011), and an Assistant Professor at Albert Einstein College of Medicine in New York (2011-2018). Having only just relocated back to Australia as an ARC Future Fellow in July 2018, a major focus of her work is on modelling the aging process using single cell bioinformatics. Jess has received several awards, including a Fulbright scholarship (2003), the Metcalf Prize for Stem Cell Research from the National Stem Cell Foundation of Australia (2017), and the LaDonne H. Shulman Award for Teaching Excellence (2017) from Albert Einstein College of Medicine.

SCALABLE BIOINFORMATICS METHODS FOR SINGLE CELL DATA

Associate Professor Joshua Ho, University of Hong Kong

Single cell RNA-seq and other high throughput technologies have revolutionised our ability to interrogate cellular heterogeneity, with broad applications in biology and medicine. Standard bioinformatics pipelines are designed to process individual data sets containing thousands of single cells. Nonetheless, data sets are increasing in size, and some biological questions can only be addressed by performing large-scale data integration. There is a need to develop scalable bioinformatics tools that can handle large data sets (e.g., with >1 million cells). Our laboratory has been developing scalable bioinformatics tools that make use of modern cloud computing technology, fast heuristic algorithms, and virtual reality visualisation to support scalable data processing, analysis, and exploration of large single cell data. In this talk, we will describe some of these tools and their applications.

Dr Joshua Ho is an Associate Professor in the School of Biomedical Sciences at the University of Hong Kong (HKU). Dr Ho completed his BSc (Hon 1, Medal) and PhD in Bioinformatics from the University of Sydney and undertook postdoctoral research at the Harvard Medical School. His research focuses on advanced bioinformatics technology, ranging from scalable single cell analytics, metagenomic data analysis, and digital healthcare technology (such as mobile health, wearable devices, and healthcare artificial intelligence). Dr Ho has over 80 publications, including first or senior-author papers in leading journals such as Nature, Genome Biology, Nucleic Acids Research and Science Signaling. Prior to joining HKU, he was the Head of Bioinformatics and Systems Medicine Laboratory at the Victor Chang Cardiac Research Institute. His research excellence has been recognized by the 2015 NSW Ministerial Award for Rising Star in Cardiovascular Research, the 2015 Australian Epigenetics Alliance's Illumina Early Career Research Award, and the 2016 Young Tall Poppy Science Award.

WORKSHOP: SINGLE CELL RNA-seq ANALYSIS

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

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.

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

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.

WORKSHOP: SINGLE CELL RNA-SEQ ANALYSIS ON THE CLOUD

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

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.

Keywords: Spark; RNA-seq; big 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.

WORKSHOP: GENE-EXPRESSION ANALYSIS WITH RNA SEQUENCE DATA USING R

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

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.

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

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.

WORKSHOP: PITFALLS AND ROADBLOCKS IN SINGLE-CELL ANALYSES

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

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.

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

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 4 – Thursday 5 December

MASS SPEC ANALYTICS

08:45-09:00	**REGISTRATION**
09:00-09:45	Challenges and questions in proteomics technologies and research Dr Mark Larance, The University of Sydney
09:45-10:30	iProFun: An integrative analysis tool to screen for Proteogenomic Functional traits Professor Pei Weng, Icahn Medical School at Mount Sinai, New York
10:30-11:00	**MORNING TEA (catered)**
11:00-11:25	Using Multiomic Technologies to Unravel Heart Failure Dr John O'Sullivan, The University of Sydney
11:25-11:50	Computational analysis for biological discovery from (phospho)proteomic data Dr Pengyi Yang, The University of Sydney
11:50-12:15	Key ingredients of a data analytical pipeline with TMT or SWATH Dr Dana Pascovici, Australian Proteome Analysis Facility, Macquarie University
12:15-13:30	**LUNCH**
13:30-15:00	WORKSHOPS jointly held with BioCAsia ***Sponsored by Children's Medical Research Institute***
Stream A	Imputation and data quality control for proteomics data Professor Pei Weng, Icahn Medical School at Mount Sinai, New York
Stream B	Computational analysis for biological discovery from (phospho)proteomic data Dr Pengyi Yang, The University of Sydney
Stream C	R and Bioconductor for Genomic Analysis Professor Martin Morgan, Roswell Park Comprehensive Cancer Center
15:00-15:30	**AFTERNOON TEA (catered)**
15:30-16:30	WORKSHOPS jointly held with BioCAsia ***Sponsored by Children's Medical Research Institute***
Stream A	Imputation and data quality control for proteomics data Professor Pei Weng, Icahn Medical School at Mount Sinai, New York
Stream B	Introduction to proteomics Dr Ben Crossett, David Maltby, Angela Connolly, Sydney Mass Spectrometry, University of Sydney
Stream C	Building a Bioconductor package Dr Peter Hickey, Walter and Eliza Hall Institute for Medical Research Dr Saskia Freytag, Harry Perkins Institute of Medical Research
16:30-17:00	Fast Forward Poster Session
17:00-18:00	Poster Q&A Session Public lecture registration (catered)
18:00-19:30	**Public Lecture: The Bright Future of Applied Statistics** **Professor Rafael Irizarry, Harvard University**

CHALLENGES AND QUESTIONS IN PROTEOMICS TECHNOLOGIES AND RESEARCH

Dr Mark Larance, The University of Sydney

Mass spectrometry when coupled with high resolution liquid chromatography is a key technology for protein analysis. We have established several workflows for the sensitive analysis of protein-protein interactions in mammalian tissues such as liver and protein abundance analysis in human blood plasma. These methods allow us to monitor the effects of various clinical interventions such as intermittent fasting to determine their mechanism of action. In this talk I will cover these methods and the integration of the datasets to derive new biological knowledge.

Mark Larance received his undergraduate degree in biochemistry from UNSW in Sydney, Australia. He was awarded a PhD in biochemistry in 2007 from UNSW, based on a collaborative project between the Garvan Institute (Prof David James) and UNSW (Prof Michael Guilhaus). During his PhD, he developed methods for the study of insulin signalling and vesicle trafficking using mass spectrometry-based proteomics.

In 2009, he moved to the University of Dundee in Scotland to take up a post-doctoral position in the laboratory of Prof. Angus Lamond. He was awarded a Royal Society of Edinburgh Personal Research Fellowship, which allowed him to become an independent investigator at the University of Dundee. During this period, he expanded his interest in starvation responses in model animals.

In 2016, he was awarded a Cancer Institute NSW Future Research Leader fellowship, which allowed him to return to Australia from Scotland to investigate how programs of intermittent fasting can assist with the prevention and treatment of cancer. He is currently a senior lecturer in the School of Life and Environmental Sciences and is in the Charles Perkins Centre (CPC), which is a multidisciplinary institute devoted to easing the global burden of obesity, diabetes, cardiovascular disease and related conditions.

IPROFUN: AN INTEGRATIVE ANALYSIS TOOL TO SCREEN FOR PROTEOGENOMIC FUNCTIONAL TRAITS

Professor Pei Weng, Icahn Medical School at Mount Sinai, New York

In this talk, I will introduce iProFun, an integrative analysis tool to screen for Proteogenomic Functional traits perturbed by DNA copy number alterations (CNAs) and methylations. The goal is to characterize functional consequences of DNA copy number and methylation alterations in tumors and to facilitate screening for cancer drivers contributing to tumor initiation and progression. Specifically, we consider three functional molecular quantitative traits: mRNA expression levels, global protein abundances, and phosphoprotein abundances. We aim to identify those genes whose CNAs and/or DNA methylations have cis-associations with either some or all three types of molecular traits. In comparison with analyzing each molecular trait separately, the joint modeling of multi-omics data enjoys several benefits: iProFun experienced enhanced power for detecting significant cis-associations shared across different omics data types; and it also achieved better accuracy in inferring cis-associations unique to certain type(s) of molecular trait(s). For example, unique associations of

CNAs/methylations to global/phospho protein abundances may imply post-translational regulations. I will show an application of iProFun on ovarian high-grade serous carcinoma tumor data from TCGA and CPTAC. The result suggests potential drug targets for ovarian cancer.

Pei Wang is a Professor in Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai. Dr. Wang obtained her B.S. in Mathematics from Peking University, China, in 2000. She then pursued her graduate study in the U.S. and received a Ph.D. in Statistics from Stanford University in 2004. Between 2004-2013, Dr. Wang served as a faculty at Fred Hutchinson Cancer Research Center and University of Washington, Seattle, WA. In Oct 2013, she joined Icahn Medical School at Mount Sinai, New York. Dr. Wang's research has been focused on developing statistical and computational methods to address scientific questions based on data from high throughput biology/genetics experiments as well as modern digital health studies. Dr. Wang and her team have developed numerous novel statistical methods for analyzing and integrating various genetic/genomic/proteomic data. In the past decade, Dr. Wang has been actively involved in the NCI funded CPTAC (Clinical Proteomic Tumor Analysis Consortium). Currently, Dr. Wang is the MPI of the national Proteomics and Genomics data analysis center of CPTAC.

USING MULTIOMIC TECHNOLOGIES TO UNRAVEL HEART FAILURE

Dr John O'Sullivan, The University of Sydney

This talk will explore novel insights into heart failure. Applying novel technologies – proteomics and metabolomics – to a unique resource – the Sydney Heart Bank – is providing novel insights into this disease. A critical component of this work is bioinformatic analysis that can merge different datasets and different data types together and uncover the key drivers of disease pathology.

The talk will also cover the emergence of a type of "stiff" heart failure called Heart Failure preserved Ejection Fraction that has taken over as the major heart failure phenotype, driven by the obesity epidemic. Situated at the Charles Perkins Centre, we are uniquely placed to make advances in this field. Combining dietary, microbiome, cardiovascular, sleep, imaging, and omic expertise and technologies to human cohorts and sophisticated model systems will enable critical new insights into this disease.

Dr John O'Sullivan is a Clinical-Academic Cardiologist at the Royal Prince Alfred Hospital and Group Leader in Cardiometabolic Disease at the Heart Research Institute and Charles Perkins Centre of the University of Sydney. John undertook his internal medicine, cardiology, and PhD training in Ireland, and then spent 4 years at Massachusetts General Hospital and Harvard Medical School studying functional genomics and metabolomics. John studies the cardiovascular consequences of obesity and related diseases such as diabetes. He is particularly interested in diabetic cardiomyopathy and cardiac metabolism in HFpEF, and the dietary-microbiome-metabolome-cardiovascular disease axis. He utilizes his clinic, murine models, cardiac tissue slice model, iPSC-cardiomyocytes, Langendorff perfusion, stable-isotope tracing, the Sydney Heart Bank, and close collaborations with colleagues at the Charles Perkins Centre to address fundamental questions in these research fields.

COMPUTATIONAL ANALYSIS FOR BIOLOGICAL DISCOVERY FROM (PHOSPHO)PROTEOMIC DATA

Dr Pengyi Yang, The University of Sydney

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. To gain the most from this talk/workshop, you are encouraged to bring your own computer and follow the hands-on demonstrations step-by-step.

Pengyi Yang is senior lecturer at the School of Mathematics and Statistics, the University of Sydney, and an ARC DECRA Fellow and a NHMRC Investigator. He heads the Computational Systems Biology group at Children's Medical Research Institute, and also heads the Computational Trans-Regulatory Biology group at Charles Perkins Centre. Pengyi has extensive experience in integrative analysis of proteomic, phosphoproteomic, transcriptomic, and epigenomic data. In particular, he has developed various specialised machine learning methods and statistical models for biologically guided analysis of multi-layered omic data.

KEY INGREDIENTS OF A DATA ANALYTICAL PIPELINE WITH TMT OR SWATH

Dr Dana Pascovici, Australian Proteome Analysis Facility, Macquarie University

As a data analysis group embedded in a proteomics facility, we look at many datasets each year, generated with various proteomics techniques. Lately the emphasis has been on SWATH and TMT, excellent techniques for discovery proteomics, each with their respective strengths. In this talk I will discuss what we see as key elements of a data analytical pipeline for such techniques. A starting point is often a decision of which technique is better suited for each project, and with it capturing in a uniform fashion as much of the experimental design as possible. Batch considerations are important for larger experiments, which are becoming more common, as well as deciding how batch effects will be handled – typically via normalisation such as IRS or other statistical methods. Analysis environments such as Perseus are fantastic for individual in-depth analyses, but automated analysis options can save time and help with reproducibility and versioning, which are important in a quality accreditation environment. In our in-house workflows we often use R based tools such as SwathXtend or TMTPrePro. And finally, a crucial background ingredient is typically a controlled, spike-in dataset which can be relied on for assessing the key methods and their subsequent fine tuning.

Dana Pascovici is a biostatistician working at the Australian Proteome Analysis Facility (APAF) based at Macquarie University, Sydney. She comes from a mathematical and computational background, having completed a bachelor degree in Mathematics and Computer Science at Dartmouth College in the US, followed by a PhD in Mathematics at MIT. For the past 15 years both in the industry and research environment, her research has been focused on generating reliable methods of interpreting and analysing data from a variety of quantitative proteomics platforms, lately emphasizing SWATH and TMT, and wherever possible incorporating them into software workflows. Areas of particular relevance to APAF's bioinformatics team have been plasma proteomics, and plant proteomics of agriculturally important species. As a lead scientist in data analysis at APAF, she has helped researchers to generate biological insights out of their proteomics data, especially in the context of complex experiments. Such work is always collaborative, and benefits from interactions with researchers, students, and the APAF team of mass spectrometry specialists and analytical chemists.

WORKSHOP: IMPUTATION AND DATA QUALITY CONTROL FOR PROTEOMICS DATA

Professor Pei Weng, Icahn Medical School at Mount Sinai, New York

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.

Key words: proteomics, imputation

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

WORKSHOP: COMPUTATIONAL ANALYSIS FOR BIOLOGICAL DISCOVERY FROM (PHOSPHO)PROTEOMIC DATA

Dr Pengyi Yang, University of Sydney

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.

Keywords: Proteomics, data mining

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.

WORKSHOP: R AND BIOCONDUCTOR FOR GENOMIC ANALYSIS

Professor Martin Morgan, Roswell Park Comprehensive Cancer Center

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.

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

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.

WORKSHOP: INTRODUCTION TO PROTEOMICS

Dr Ben Crossett, Sydney Mass Spectrometry, University of Sydney
David Maltby, Sydney Mass Spectrometry, University of Sydney
Angela Connolly, Sydney Mass Spectrometry, University of Sydney

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.

Key words: mass spectrometry, proteomics

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.

WORKSHOP: BUILDING A BIOCONDUCTOR PACKAGE

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

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.

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

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.

AMSI 19

BIOINFO SUMMER

A SYMPOSIUM IN BIOINFORMATICS

CHARLES PERKINS CENTRE
THE UNIVERSITY OF SYDNEY
2–6 DECEMBER

POSTER ABSTRACTS

AMSI AUSTRALIAN MATHEMATICAL SCIENCES INSTITUTE
Australian Government Department of Education
THE UNIVERSITY OF SYDNEY

A PORTABLE BIOINFORMATICS PIPELINE FOR THE ANALYSIS OF LARGE-SCALE RNA-SEQ DATASETS

Hamish Mundell, The University of Sydney

In recent years, Next Generation Sequencing has led to the generation of several very large datasets. These datasets require complex processing steps to produce a meaningful output and thereby enable powerful downstream analysis. A lot of these data sources are publicly available, but due to their inherent processing complexity, they are an untapped resource. There is the potential to conduct large-scale studies which bring together multiple, related datasets from a number of studies: For example, WGS data transformed into RNA-Seq data, which can be used to later train models which distinguish Parkinson's Disease (PD) patients from healthy controls. We have developed several cutting edge bioinformatic pipelines, which seek to automate these processing steps in a computationally fast and efficient manner, while also being scalable and reusable across multiple platforms and datasets. Here we introduce one such bioinformatics pipeline, developed to enable the retrieval, processing and transcriptomic analysis of ~4000 RNA Sequencing samples from multiple different studies. We use the programs Jenkins, Cromwell and Kallisto in the development of this pipeline and demonstrate its application to the datasets of interest.

LOSS OF FUNCTIONAL GASTRIC GENES IN THE MONOTREME LINEAGE

Natasha Bradley, The University of Adelaide

Monotremes (platypus and echidna) are the oldest surviving mammalian linage, having diverged from therian mammals 190 million years ago. Monotremes have unique biology, combining elements of reptiles, birds and mammals. Interestingly, the stomach of monotremes is very small and the gastric juice of the monotreme stomach has a neutral pH of 6.2-7.4, while other mammals are acidic. Additionally, monotremes are missing all glands in the stomach, except for the Brunner's glands. This renders the monotreme stomach non-functional (in comparison to other mammalian stomachs) and appears an elongated oesophagus.

There was no molecular evidence of these observations until the publishing of the platypus genome in 2008, where various genes involved in stomach function were found to be missing or pseudogenised. These genes were involved in gastric acid secretion and protein degradation. Through searching an unpublished draft echidna genome and improved platypus genome, we have found the loss of these genes occurred not only in the platypus, but also in the echidna. This indicates that the loss of gastric genes occurred in the monotreme ancestor, giving insight into their evolution.

AN INTEGRATIVE APPROACH TO TISSUE-SPECIFIC EFFECTS OF MICRO-RNA REGULATORY NETWORKS

Tânia Marques, Faculty of Sciences / University of Lisbon

Dr Nham Tran, University of Technology Sydney

Dr Margarida Gama-Carvalho, Faculty of Sciences / University of Lisbon

miRNAs are small noncoding RNAs with role in post-transcriptional regulation of gene expression. Even though the basic mechanisms for miRNA action have been described, we are still unable to efficiently predict their impact on cellular function. When predicting targets for a miRNA, the context in which it is expressed needs to be accounted for. The aim of this work is to comprehensively understand the interaction dynamics between miRNAs and their target transcriptome across different tissues.

Paired miRNA and mRNA normalized count data corresponding to normal samples of seven tissues present in TCGA were downloaded. The samples were clustered to understand if any outliers were present. The profile of miRNA expression was assessed, and categories to include the miRNA or transcripts were created according to their level of expression and tissue specificity. The interactions of miRNAs and their targets were investigated through correlation analysis and compared across tissues. The clustering analysis revealed that, overall, the tissues cluster well together. The categories and levels of expression of miRNA do not reflect a higher number of negatively correlated transcripts (-0.5). The results of this work provide a deeper understanding of the regulatory networks governing gene expression regulation and allow further explorations of miRNA action.

DEEP ANALYSIS OF MIRNA AND ISOMIR EXPRESSION ON A SINGLE CELL LEVEL

Christopher Smith, University of Technology Sydney

Professor Gyorgy Hutvagner, University of Technology Sydney

miRNAs are small non-coding RNAs that play a role in the post-transcriptional regulation of genes. Studies have shown that miRNAs from the same precursor can vary in their exact sequence due to cellular mechanisms such as alternative drosha/dicer trimming or untemplated nucleotide additions, which may affect the miRNAs stability or target genes. Pioneering studies have already used single cell RNA-seq to highlight the complex heterogeneity of gene expression in cells or tissues previously assumed to be homogenous, and cancer research is beginning to focus on using this knowledge to develop novel treatments that can target different cell populations in tumors. Despite recent technological developments enabling single cell small RNA sequencing, very few studies have investigated how miRNAs and isomiRs are expressed on single cell level, and what role they may have in contributing to cell heterogeneity in healthy or diseased tissue remains unanswered. In this study we apply bioinformatics to published single cell small-RNA datasets to investigate miRNA and isomiR expression on a single cell level and evaluate the potential for using isomiRs as predictors of cell identity.

SINGLE CELL RNA-SEQ ANALYSIS REVEALS THE HETEROGENEITY OF THE COLONIC MESENCHYMAL CELLS IN INFLAMMATORY BOWEL DISEASE

Junwei Wang, The University of Adelaide

Dr Stephen Pederson, The University of Adelaide

Inflammatory bowel disease (IBD), with ulcerative colitis and Crohn's disease as two major forms, is characterized by chronic relapsing intestinal inflammation, and has been a worldwide healthcare problem with increasing incidence and prevalence. Its specific etiology is so far poorly understood. Mesenchymal cells of the intestinal lamina propria play important roles in immune homeostasis, and epithelial barrier maintenance. Their function was impaired in IBD through poorly defined pathways. Colonic mesenchymal cells are highly heterogeneous with overlapped marker genes, which prevented the study of their cell-type specific functions and attributes to IBD. We used a public colonic mesenchymal scRNA-seq dataset, to reveal the heterogeneity of colonic mesenchymal cells in IBD. We observed differential gene expression between stromal cells from healthy and colitis. Stromal cells with 5 subtypes from the mouse dataset, and 3 subtypes from the human data was detected. Also, 2 subtypes from mouse data, and 1 subtype from human data were detected to be the colitis-associated, with the rest shown to be healthy-associated. We also identified specific marker genes for each of these stromal subtypes. Our finding will assist further investigation of specific attributes of mesenchymal stromal subtypes, and enhancing immune monitoring of existing and new therapies in IBD.

TRANSCRIPTOMICS IN NEURODEGENERATIVE DISEASE

Kosar Hooshmand, The University of Sydney

Neurodegenerative Disease (ND) is an umbrella term for a range of conditions which primarily affect the neurons of the human brain. So far there are no effective biomarkers for monitoring NDs and their progression. A multi modal investigative approach is required for identification of specific clinically relevant biomarkers, necessary for therapeutic advancement in NDs.To identify the genetic/biological processes, which ultimately lead to the development of target therapeutics in NDs we are developing a platform which combines and analyses multi RNA-Seq data. Large downloads of multiple high-content data and statistical analysis models for these large datasets will enhance knowledge of core differentially expressed molecules and similarly perturbed pathways for their potential as biomarkers of NDs both in general and specific to different disorders, as well as appropriate therapeutic targets. Our group has dedicated computing resources, tools and workflows(Workflow Description Language (WDL) & the Common Workflow Language (CWL)) to process and analyse large and complex biological datasets (RNA sequence analysis) using machine learning/artificial intelligence and feature selection methods for identification of genetic or structural variants, expression changes and distorted alternative splicing events in relevant brain regions that can serve as potential biomarkers for the diagnosis and prognosis of various diseases including NDs.

INTROME: IDENTIFYING SPLICE-ALTERING VARIANTS AS DRIVERS OF HIGH-RISK PAEDIATRIC CANCER

Patricia Sullivan, Children's Cancer Institute

Genetic variants that affect pre-mRNA splicing can have a substantial impact on the resulting protein. Identifying and predicting a variant's impact on splicing is challenging, and current bioinformatic methodologies frequently miss these potentially medically-relevant variants.

To address this area of need, we have created and optimised Introme, a bioinformatic tool designed to identify and predict the impact of splice-altering variants. Introme uses machine learning (C5.0) to integrate predictions from multiple splice scoring tools, evaluating the likelihood of a variant to impact splicing. We applied Introme to 250 paediatric cancer patients, analysing a subset of known cancer genes in the germline and tumour WGS results.

We curated a list of 802 splice-altering variants from 150 papers to form the training and validation set for machine learning. Introme achieved the best performance (AUC: 0.96) of the tools evaluated, followed by SpliceAI (AUC: 0.93) and MMSplice (AUC: 0.81). Introme's predictions have led to the identification of 140 RNA-seq validated splice-altering variants in patients with paediatric cancer.

The application of Introme to patient sequencing data uncovers aberrations that were missed by previous analysis methods. Detecting these splice-altering variants has aided the identification of medically-relevant variants and facilitates the recommendation of personalised treatment options.

FINDING ALTERNATIVE POLYADENYLATION SIGNATURES AS CANCER BIOMARKERS

Nitika Kandhari, Monash University

Dr Paul Harrison, Monash University

Associate Professor Kaylene Simpson, The University of Melbourne

Associate Professor David Powell, Monash University

Associate Professor Traude Beilharz, Monash University

The output of cellular transcription is diversified by alternative RNA processing. For example, in addition to differential splicing, 70% of mammalian genes undergo Alternative Polyadenylation (APA). This changes the architecture of 3'-UnTranslated Regions (UTRs) and associated post-transcriptional regulatory control of mRNA fate. Short 3'UTRs are generally associated with de-differentiated proliferative cells (e.g. stem cells) whereas longer 3'-UTRs associate with more complex regulation and cellular specialisation. The literature suggests that ~91% APA genes switch to shorter mRNA isoforms in tumour cells. Our study aims to detect a signature of APA changes that are specific to triple-negative breast cancer (TNBC) that could be applied as a novel prognostic biomarker in early-stage breast cancer. Using bioinformatic analyses of 3'-focused RNA-seq approaches we studied the landscape of transcription and APA in three cancer cell lines in response to loss of PCF11, a core regulator of 3'-end formation. This shows a conservation of an expression and processing response to loss of 3'-end processing machinery. In addition to gene expression changes, we identify 3'-end "shifted" genes that are common to all 3 cell lines. We will present our current work around the idea that systematic lengthening of 3'UTR might normalise deregulated gene expression in cancer.

INVESTIGATING CHEMORESISTANCE MECHANISMS IN NEUROBLASTOMA USING A LONGITUDINAL CISPLATIN RESISTANCE MODEL

Janith Seneviratne, The University of New South Wales
Dr Anushree Balachandran, Children's Cancer Institute
Mrs Claudia Flemming, Children's Cancer Institute
Dr Marion Le Grand, Children's Cancer Institute
Professor Maria Kavallaris, Children's Cancer Institute
Professor Glenn Marshall, Children's Cancer Institute
Dr Belamy Cheung, Children's Cancer Institute
Dr Daniel Carter, Children's Cancer Institute

Neuroblastoma is the most common extracranial solid tumour in children. 50% of children with high-risk neuroblastoma relapse, often due to drug resistance following induction chemotherapy.

To investigate drivers of chemoresistance in neuroblastoma, we established a longitudinal drug resistance model. IMR-32 cells were pulsed with increasing concentrations of cisplatin over 13 months. Three increasingly resistant cell lines were established at separate time points. We used the 10X Genomics Chromium platform to transcriptionally profile single cells across this resistance trajectory.

We examined gene expression signatures of individual cells to 1) distinguish unique transcriptional cell states, 2) reconstruct transcriptional trajectories, and 3) infer changes in copy number variations (CNVs).

We identified three transcriptionally distinct clusters with varying abundances of cells from all four cell lines. Depending on the cluster assessed, we identified depletion or expansion of transcriptomic phenotypes in response to cisplatin, suggesting that transcriptional plasticity contributes to drug resistance. Through trajectory analysis we observed transitions in cytoskeletal genes with increasing cisplatin exposure. We further linked these changes in gene expression to inferred segmental CNVs found at Chr1p36.

Taken together our work suggests that upon cisplatin exposure, neuroblastoma cells become resistant using a combination of both genetic (innate) and non-genetic (acquired) mechanisms.

SOMETHING'S FISHY ABOUT MY DATA: CHALLENGES WITH REPLICABILITY IN RNA-SEQ

Lachlan Baer, The University of Adelaide

Replicability of results obtained from comparable RNA-seq datasets is largely sought after, however, is not consistently achieved. An initial comparison between two analogous RNA-seq datasets from zebrafish brain tissue identified that substantial batch effects were involved. Further investigation into differences seen between multiple datasets determined a number of factors to contribute to overall variability. Possible sources of variation may occur during the sample and library preparation stages prior to RNA-seq. This suggests that the choice in preceding techniques may have an impact on experimental outcomes, with the potential to obscure particular observations when biological signals

are subtle. An attempt to remove unwanted variation was limited in success. However, consistent results between datasets when analysed for global trends in expression patterns suggests that technique-dependent biological interpretations may still be possible.

SINGLE-CELL RNA-SEQ ANALYSIS TO EXPLORE BONE-MARROW IMMUNE LANDSCAPE

Gunjan Dixit, Australian National University

Bone marrow (BM) contains multiple immune cell subsets with critical functions and is considered an immune regulatory organ. It contains osteoclasts and immune cells fundamentally involved in physiological and pathological bone remodelling. In autoimmune diseases, inflammation can impair the BM niche, disturb hematopoietic and immune development, and induce osteoporosis. Specific cytokines exhibit pleiotropic effects on the immune system, and their discovery in the regulation of survival, differentiation and propagation of activated T cells paved the path for its direct clinical implications in immunotherapy. This project addresses the fundamental question of how low-dose of CytokineX modulates BM immune landscape by comprehensively mapping the therapy-induced changes using single-cell technologies. I analyzed the scRNA-seq data obtained from BM of CD45 cells of mice to identify different immune cell types and compared their expression across four experimental conditions- a control (sham), control treated with CytokineX (Sham+Treatment), ovariectomy-induced osteoporosis (OVX) and OVX treated with CytokineX (OVX+Treatment). The analysis pipeline included quality control, normalization, data integration, dimensionality reduction, clustering and downstream processing. Pairwise differential expression analysis confirmed the identified cell-types and their assignment into clusters. Trajectory analysis using RNA velocity revealed the cellular dynamics and the lineage of sub-populations.

PROFILING GENE DYSREGULATION CAUSING INHERITED PERIPHERAL NEUROPATHY BY CHARACTERISING OF TOPOLOGICAL ASSOCIATED DOMAINS (TADS) IN 3D GENOME

Dr Kaitao, Lai, The University of Sydney

Mr Anthony N.Cutrupi, The University of Sydney

Dr Gonzalo Perez-Siles, The University of Sydney

Professor Garth A. Nicholson, The University of Sydney

Professor Marina L. Kennerson, The University of Sydney

Inherited peripheral neuropathies (IPNs) cause degeneration of peripheral nerves with more than 100 causative genes reported to date. Our studies have demonstrated that how SV can impact on the 3D genome of IPN patients.

We have identified a 1.35 Mb insertion of chromosome 7q36.3 causing gene dysregulation in a large multi-generation family linked to the distal hereditary motor neuropathy DHMN1 locus. We suppose that the DHMN1 complex insertion disrupts genomic organisation leading to gene dysregulation and subsequent axonal degeneration.

We performed high-throughput chromosome conformation capture (Hi-C) analysis of disease-associated mutations and chromosomal rearrangements using DHMN1 induced pluripotent stem cell derived motor neurons (iPSC-MNs) to investigate the mechanism of chromatin interaction underlying the gene dysregulation.

In our works, the 2D contact heatmaps have been generated and the topologically associated domain (TAD) data have been predicted in the view of 3D genome. The neo-TAD on the DHMN1 locus has been identified, which presents duplication and insertion in 3D genome level and may suggest overlapping duplications that extend over the next boundary into the neighbouring regulatory domain. In addition, the 3D representations of the above 2D heatmaps have been constructed, which indicate that the DHMN1 complex insertion has altered the 3D chromatin loops.

THE DYNAMIC GENOME BEHIND THE EMERGENCE OF RECENT OCTOPOD NOVELTIES

Brooke Whitelaw, James Cook University

Cephalopods are characterized by many organismal novelties. To reveal the genomic correlates of organismal novelties, we conducted a comparative study of three octopod genomes. Among the species examined is a member of the blue ringed octopus genus (Hapalochlaena) the only known octopods to store large quantities of the potent neurotoxin tetrodotoxin (TTX) within their tissues and venom gland. We present the first genome of a member of this genus, the southern blue-ringed octopus (Hapalochlaena maculosa) and reveal highly dynamic genome evolution at both non-coding and coding organizational levels. We demonstrate expansions of zinc finger and cadherin gene families associated with neural functions/tissues in both H. maculosa and C. minor are congruent with the previously observations in O. bimaculoides, suggesting an octopod specific trait. Examination of tissue specific genes in the posterior salivary/venom gland (PSG) revealed putative venom proteins, serine proteases dominate expression in O. bimaculoides and C. minor, while representing a minor component in H. maculosa. Voltage-gated sodium channels (Nav) in H. maculosa contain a resistance mutation previously documented in pufferfish and garter snakes to confer 10-15 fold resistance to TTX. No known resistance mutations were identified in either O. bimaculoides or C. minor .

WHAT EXACTLY DO WE KNOW ABOUT FUNCTIONAL CONSTRAINT ON RENAL GENES ASSOCIATED WITH ESRF?

Hope Tanudisastro, The University of Sydney
Mahnoor Bakhtiar, The University of Sydney
Dr Yuan Min Wang, The University of Sydney
Dr Geoff Zhang, The University of Sydney
Dr Hugh McCarthy, The University of Sydney
Professor Stephen Alexander, The University of Sydney

Large-scale exome sequencing data have allowed for the interrogation of low frequency variations and their predicted pathogenicity. It is hypothesised that, in certain genes, selective constraint diminishes observed functional variation.

Using data from the Exome Aggregation Consortium of 60,706 patients, we examined functional constraint on genes associated with end stage renal failure (ESRF) by comparing the frequency of synonymous, missense, and loss-of-function (LoF) mutations against their respective selection-neutral expected values, taking into account gene length, read depth, and local sequence context. ESRF-associated genes were identified from clinical data on paediatric patients in Boston Children's Hospital and the Children's Hospital at Westmead. We compared these values across genes identified by transcriptomic analysis to be highly tissue-enriched in kidneys and those associated with chronic kidney disease (CKD) by genome-wide association study.

Stronger negative selection was observed in ESRF-associated genes than in kidney-expressed genes across LoF mutations ($p<6.6e-07$). Amongst missense mutations, autosomal dominant ESRF-associated genes are under more selective constraint than kidney-expressed genes ($p<0.004$). CKD-associated, ESRF-associated, and kidney-expressed gene sets had similar z-score distributions for synonymous mutations (p = n.s.). Selective pressure was measured most strongly across LoF mutations while genes highly enriched for renal tissue experience relatively minimal negative selection.

MULTI MODAL MACHINE LEARNING TO INFORM BETTER DIAGNOSES, PROGRESSION PROGNOSES AND CLINICAL TRIALS FOR PARKINSON'S DISEASE

Michael Allwright, The University of Sydney

I am employing cutting edge machine learning/Artificial Intelligence algorithms and data management processes to interrogate and bring together multi-modal datasets relating to Parkinson's Disease (PD). Through the integration of these data sources and the application of machine learning methods, I am seeking to develop improved biomarkers thus enabling an improved prediction of disease diagnosis, disease sub-type and disease progression.

These biomarkers will in turn enable patients to receive better treatment plans, bespoke to their specific disease sub-type and rate of progression. It will also enable improved efficiency in clinical trials, due to improved targeting of patient cohorts by disease stage and type.

The data sources considered are:

Publicly available resources such as Michael J Fox/PPMI data and next generation sequencing data as well as clinical data available at Sydney University's Brain and Mind Institute;

Medical Imaging Data, Whole Genome Sequencing Data, Lipidomics, Blood and Metabolonic Data.

The goal is to develop an automated, multi-modal Big Data and machine learning pipeline using the latest technologies (R, Kallisto, Cromwell etc.), which is reusable and scalable and which can be implemented on any relevant data sources to achieve the goals highlighted above.

FORMATION OF BOUNDARIES IN THE DEVELOPING EMBRYO

Bin Wang, Monash University

The heart is one of the most crucial organs throughout life. Its development arises from the embryonic mesoderm and this process is strictly regulated by tightly controlled, spatio-temporal gene expression. Currently, the specific expression network which governs cardiac expression boundaries remains largely cryptic. Using the Tomo-seq database generated by Junker et al. in 2014, of the developing Danio rerio embryo, gene expression can be specifically attributed to 3D spatial regions. By combining known cardiac markers with a powerful computational approach, it is possible to reconstruct detailed gene regulatory networks of the developing heart. Clustering methods can divide similarly expressed genes into groups whereby gaining new biological knowledge following annotation. The hierarchical clustering algorithm is applied to identify and characterise unannotated genes in the LPM domain along the left to right, and anterior to posterior axes. In this way, a comprehensive series of genes active in the LPM territory can be used to inform the reconstruction of a computational GRN, providing new insights into the complex development of cardiac boundaries. In the future, further characterisation of these candidate genes controlling cardiac boundary networks may provide useful insights in clinical intervention in congenital and adult heart disease.

THE ROLE OF KINASES IN THE DIFFERENTIATION OF BONE MARROW STROMAL CELLS INTO OSTEOBLASTS: A SYSTEMATIC ANALYSIS BY KNOCKDOWN

Angelita Liang The University of New South Wales

Kinases may play positive or negative roles in the transduction of key signalling pathways which regulate the differentiation of human bone marrow stromal cells (hBMSCs) into osteoblasts. To discover novel kinases which regulate hBMSC differentiation into osteoblasts, we analysed the data from a systematic RNA interference knockdown of 719 genes including 393 human protein kinases in the hMSC-TERT4 cell line induced to differentiate over 6 days using alkaline phosphatase as a proxy for the extent of differentiation. We have then cross-referenced the putative regulators we found with their mRNA expression profiles over 12 days as measured by time-series RNA Sequencing. By combining knockdown data with large-scale CRISPR knockout data that indicated the essentiality of a gene for cell viability, we identified the genes JAK1,GSK3B,ERBB2 and EPHA3 to be putative positive regulators and PRKCE,PTK7,TEC,PRPS1,PRPS2,PRPSAP1 and PRPSAP2 to be putative negative regulators. We found genes encoding members of the MAPK,MAP2K,MAP3K,MAP4K and NIMA-related kinase families comprising the positive regulator hits, and genes encoding members of the Casein kinase, Protein kinase G, and RIPK families in the negative regulator hits...

FROM DPP4 SUBSTRATES TO BEYOND

Robert Qiao, Flinders University

From type II diabetics, rheumatoid arthritis to Alzheimer, dementia even cancers, chronic diseases create severe social-economical burdens on the healthcare system and patients' welfare are compromised in the wake of prolonged drug usage in many cases.

With ever-increasing omics data becomes available, we now have a unique opportunity to study the global interactions and correlation network of a given enzyme inhibitor therapy. This study attempted to reveal the global interaction network for dipeptidyl dipeptidase 4 (DPP4), and to further explore the possible long-term adverse effects due to prolonged inhibition of DPP4 activity, given that DPP4 has emerged as a new therapeutic target for the treatment of type II diabetics in clinics. The result of this study has obtained strong convergence from both data mining approach and model predictions based on machine learnings, which has depicted a much broader and diverse interaction network then literature suggests.

COMPOSITE SELECTION SIGNALS IN PUREBRED DOGS

Victor Wei Tse Hsu, The University of Sydney

Composite selection signals (CSS) have demonstrated that the locus or interesting regions can be localized in multi-breed populations. Here, we investigated the application of CSS to canine SNP data to assess previously known signatures or identify novel regions from various purebred dogs breed comparisons. We tested the use of CSS to reveal regions associated with canine disease, using lymphoma susceptibility as an example. To gain a more comprehensive insight into lymphoma predisposition, we performed a study using the CSS to analyze selected regions for potential impact on lymphoma incidence. 364 Bullmastiffs were used as a target group with a number of comparative reference groups derived from single or combined breed data. A SNP dataset of gray wolves, previously reported in a European study, was used as a source of ancestral alleles. Using the ancestral or convergent sweeps, clusters of signatures of selection were detected at 101 regions on nine canine autosomes. A gene ontology and pathway analysis of genes in regions identified by CSS revealed 89 candidate genes with enrichment for lymphoma-associated ontologies. The most significant signals were related to the regulation of lymphocyte migration. The CSS is a useful tool for cross-species for identifying potential candidate genes under selection.

BIOINFORMATIC ANALYSIS OF PHOSPHORYLATION-BASED CELLULAR SIGNALLING IN A MODEL OF EPILEPTOGENESIS

Mariella Hurtado Silva, Children's Medical Research Institute
Miss Annika Mayer, The University of Bonn
Professor Susanne Schoch, The University of Bonn
Professor Dirk Dietrich, The University of Bonn
Dr Ashley J. Waardenberg, James Cook University
Dr Mark E. Graham, Children's Medical Research Institute

The application of bioinformatics approaches to various types of â€˜omics data allows greater understanding of biological pathways that are relevant to neurological disease mechanisms. We obtained proteome and phosphoproteome data from a model of temporal lobe epilepsy at 4-hour and 24-hour after the stimuli (an injection with pilocarpine to induce status epilepticus for 30 min). The raw data contained information ono > 40,000 phosphopeptides and was normalized, filtered and reformatted prior to formal analysis. The analysis of the protein phosphorylation data provided information on the most relevant signalling pathways via gene ontology enrichment analysis. The gene ontology terms indicated that phospho-signalling activates transcription factors that respond to

strong stimuli and promote neurogenesis. To obtain information on the protein kinases responsible for phospho-signalling, we used KinSwing. KinSwing is a recently developed tool that enables the prediction of protein kinase activity. A number of regulated protein kinases were predicted to be important for each time point and indicate the activation of particular pathways. These bioinformatic analyses enabled insights into how the phosphoproteome is perturbed in the early stages of epileptogenesis.

PHENOTYPE-DRIVEN ANALYSIS OF HUMAN PHOSPHOPROTEOMES

Elise Needham, The University of Sydney

Prioritising the most important phosphosites is a major challenge in phosphoproteomics studies. Often thousands of phosphosites may be regulated and generally only a certain subset of these may drive the biological phenotype of interest. We have developed a method to focus human phosphoproteomics data analysis towards the biologically relevant phenotype. We take advantage of the variability between different humans and directly link this to phenotype. Rather than averaging across humans to compare means for particular treatment groups, we used careful experimental design involving repeated measures within individual humans. We measured the phenotype of interest, in this case glucose uptake, at every time point a sample was taken for phosphoproteomics. With both phenotype and phosphorylation measures across a cohort of humans over 4 different perturbations, we correlated phosphorylation profiles to the phenotype profile. The 152 strongly correlating phosphosites ($r > 0.7$, adjusted p-value < 0.05) out of all 1689 regulated phosphosites (Fold change > 1.5, adjusted p-value < 0.05) were enriched in known glucose uptake regulating phosphosites. The glucose uptake-correlating phosphosites also included novel phosphosites on known glucose uptake-regulating proteins, and potential new regulators. This method extracts the most relevant phosphosites to a phenotype to lead to biological insights.

LEVEL DEPENDENT QBD MODELS FOR THE EVOLUTION OF A FAMILY OF GENE DUPLICATES

Jiahao Diao, University of Tasmania

In our paper, we consider a detailed model with multi-dimensional state-space which consists of binary matrices where rows of a matrix correspond to genes, columns correspond to functions, and the ijth entries record whether or not gene i performs function j. The large state space of this model makes it unsuitable for numerical analysis, but by considering the behaviour of this detailed model we can test the suitability of two alternative models with more tractable state-spaces.

Next, we consider the model proposed in [Teufel 2014], a quasi-birth-and-death process (QBD) with two-dimensional states (n, m). (n, m) records the number n = 1, 2, . . . of genes in the family, and the number m = 0,1,...,n of redundant genes (permitted to be lost). We contrast this to a level-dependent QBD with three-dimensional states (n, m, k) that record additional information k = 1, . . . , K which affects the transition rates.

We show that two-dimensional states (n, m) model is insufficient for meaningful analysis, while the three-dimensional states (n,m,k) model is able to capture the qualitative behaviour of the detailed

model. We illustrate the fit between the level-dependent QBD and the original, detailed model, with numerical examples.

TREE SHAPE STATISTICS OF TREES GENERATED USING PHASE TYPE DISTRIBUTED TIMES TO SPECIATION

Albert Christian Soewongsono, University of Tasmania

Associate Professor Barbara Holland, University of Tasmania

Dr Malgorzata O'Reilly, University of Tasmania

This talk will some present preliminary findings in examining tree balance statistics for trees generated using a Coxian Phase Type (PH) distribution of waiting times until speciation. Some earlier results (Hagen et al, 2015) have tried to fit a model that matches with empirical tree data by analysing their tree balance statistics. One of those models is by applying a speciation rate that decreases over species age. This was done by imposing Weibull distribution with shape parameter less than one for speciation time. The biological motivation for using the Weibull was the assumption that a species can be viewed as a collection of i.i.d large populations. The simulation done using that model suggest results that match thousands of empirical trees in terms of their balance. However, viewing those sub-populations as being i.i.d may not be biologically reasonable. Here, we will be using PH distribution, specifically Coxian PH distribution to analyse the problem. The justification in using PH is due to its denseness in the field of all positive-valued distribution and using Coxian PH because every acyclic PH distributions have Coxian PH representation. The early observations using simulations with PH type show a prospective direction towards fitting to empirical tree data.

AN OMICS TRIANGLE: A CASE STUDY OF TRNA GUANINE AND INOSINE-N1-METHYLTRANSFERASE TRM5 IN ARABIDOPSIS THALIANA TO INVESTIGATE THE IMPORTANCE OF TRNA MODIFICATIONS USING TRNA-SEQ, RNA-SEQ, AND PROTEOMICS

Pei Qin (Sabrina) Ng, The University of Adelaide

Transfer RNAs (tRNAs) are critical players in messenger RNA (mRNA) decoding and are often chemically modified. It is essential to understand the consequences of losing tRNA modification by studying tRNA base changes at single-nucleotide resolution. Bioinformatics analysis of tRNA-seq has a higher sensitivity towards less abundant tRNA isotypes. Hence, the combination of tRNA-seq, RNA-seq, and proteomics data analysis enable us to study the molecular consequences of tRNA modification loss. Here, we present a case study of tRNA modifications N1-methylguanosine (m1G) and N1-methylinosine (m1I) at tRNA anticodon loop position 37 (tRNA37), which is essential to maintain translational fidelity by preventing translational frameshift. In Arabidopsis thaliana, AtTRM5, a tRNA Guanine and Inosine-N1-methyltransferase, modifies tRNA37. We show that Attrm5 mutant plants lose m1G and m1I at position 37 in tRNA-Ala and tRNA-Asp using tRNA-seq. Attrm5 mutant plants have overall slower growth and reduced primary root length. Hence, we performed RNA-seq and proteomics data analysis on both wild type and Attrm5 mutant Arabidopsis thaliana plants to examine the changes in gene expression and protein levels. In summary, our triomics approach allows

us to gain a greater understanding of how tRNA modification loss affects gene and protein expression, thus impacting plant growth and development.

INTEGRATION OF 'OMICS TECHNIQUES IDENTIFIES EXTENSIVE MITOCHONDRIAL BIOGENESIS AFTER ENDURANCE TRAINING OF HUMAN SKELETAL MUSCLE.

Dr Nikeisha Caruana, The University of Melbourne

In addition to generating the bulk of cellular energy, mitochondria direct a vast array of biological functions essential for cellular homeostasis. A long-standing question in biology concerns the biogenesis of mitochondria and its regulation in response to stress and the metabolic needs of the cellular environment, with exercise representing a major challenge to both these pathways. In order to further demonstrate the effects exercise has on the mitochondria, ten participants underwent three different training volume phases over 12 weeks. Tissue biopsies were taken prior to commencing the study and after each phase, with each mitochondrial isolated proteome analysed by label-free quantitative mass-spectrometry. Proteomics was then integrated with RNA sequencing from muscle biopsies in order to identify trends within both datasets. We observed extensive mitochondrial biogenesis in response to changing volumes of exercise training. While this was met with an overall increase in oxidative capacity, mitochondria underwent extensive remodelling of energetic pathways. Cessation of high-volume exercise reversed some, but not all of these changes. Our findings suggest that training volume is an important determinant of changes in mitochondrial content and function and is a useful model to help to further our understanding of the fundamental mechanisms of mitochondrial biogenesis.

RNA-SEQ ANALYSIS IN A ZEBRAFISH MODEL OF ALZHEIMER'S DISEASE HIGHLIGHTS THE IMPORTANCE OF IRON HOMEOSTASIS

Nhi Hin, The University of Adelaide

Analysing gene expression data from diseased and normal brain tissue has been valuable for exploring molecular mechanisms contributing to Alzheimer's disease and how these diverge from normal aging. Our laboratory has used gene editing technologies to introduce familial Alzheimer's-like mutations into zebrafish, followed by RNA-sequencing of wild-type and mutant brains at young and old age under both normal and low-oxygen conditions in a full-factorial design. This design has offered us a unique opportunity to study the molecular basis of the disease in its early stages in the young brains, and augment our knowledge with how aging and brain oxygen levels relate to disease progression. This poster describes exploratory analyses from RNA-seq data from the brains of these zebrafish and how they allow us to explore broad-scale disruptions in molecular processes, in addition to more detailed analyses testing whether processes hypothesised to be important in Alzheimer's disease (iron homeostasis in particular) were disrupted at the regulatory level. An important theme of the poster is the importance of data visualisation in facilitating hypothesis generation and communicating findings in an accessible way.

DRUGS MODULATING STOCHASTIC GENE EXPRESSION AFFECT THE ERYTHROID DIFFERENTIATION PROCESS

Dr Anissa Guillemin, The University of Melbourne

To understand how a metazoan cell takes the decision to differentiate, we study the role of stochastic gene expression during the erythroid differentiation process. It has been settled that SGE participate in decision making at the single cell level. However, experimental evidence of the relation between is still lacking. Using single cell transcriptomic analyses on avian erythropoiesis, we selected 3 drugs able to modulate the level of SGE.

We then assessed if these drugs that modulate SGE can also affect the differentiation process. We show that drugs reducing the SGE amount, significantly decreased the percent of differentiated cells and inversely.

We used a mathematical model to estimate which parameters were modified by drug treatment. We observed that among the affected parameters of the model, the rate of differentiated cells remains the parameter the most strongly affected by all drugs, supporting the previous results.

Therefore, using single-cell analyses and modeling, we provide the first evidence for a positive relation between SGE level and cell differentiation, leading to a new potential way to control this process.

TRACKING LEUKOCYTES IN INTRAVITAL TIME LAPSE IMAGES USING 3D CELL ASSOCIATION LEARNING NETWORK

Marzieh Rahmani Moghadam, La Trobe University

Leukocytes are key cellular elements of innate immune system in all vertebrates, which play a crucial role in defending organism against invading pathogens. Tracking these highly migratory and amorphous cells in in vivo models such as zebrafish embryos is a challenging task in cellular immunology. As temporal and special analysis of these imaging datasets by human operator is quite laborious, developing automated cell tracking method is highly in demand. Despite the remarkable advances in the cell detection, this field still lacks powerful algorithms to accurately associate the detected cell across time frames. The cell association challenge is mostly related to the amorphous nature of cells, and their complicated motion profile through their migratory paths. To tackle the cell association challenge, we proposed a novel deep-learning-based object linkage method. For this aim, we trained our proposed 3D cell association learning network (3D-CALN) with enough manually labelled paired 3D images of single fluorescent zebrafish's neutrophils from every two consecutive frames. A comparison of our tracking accuracy with other available tracking algorithms shows that our approach performs well in relation to addressing cell tracking problems.

AMSI BIOINFOSUMMER 2019 PUBLIC LECTURE
THE BRIGHT FUTURE OF APPLIED STATISTICS

Professor Rafael Irizarry, Harvard University

Statistics has been at the center of many exciting accomplishments of the 21st century, with applied statistics being widely used across several industries and by policy makers. In academia, the number of statisticians becoming leaders in other fields like environmental sciences, human genetics, genomics, and social sciences continues to grow. The unprecedented advances in digital technology during the second half of the 20th century has produced a measurement revolution that is transforming the world. Many areas of science are now being driven by new measurement technologies and many insights are being made by discovery-driven, as opposed to hypothesis-driven, experiments. The current scientific era is defined by its dependence on data and the statistical methods and concepts developed during the 20th century provide an incomparable toolbox to help tackle current challenges. In this talk I will give several specific examples including some from own research in genomics and estimating the effects of Hurricane María in Puerto Rico.

Rafael Irizarry received his Bachelor's in Mathematics in 1993 from the University of Puerto Rico and went on to receive a Ph.D. in Statistics in 1998 from the University of California, Berkeley. His thesis work was on Statistical Models for Music Sound Signals. He joined the faculty of the Department of Biostatistics in the Johns Hopkins Bloomberg School of Public Health in 1998 and was promoted to Professor in 2007. He is now Professor of Biostatistics and Computational Biology at the Dana-Farber Cancer Institute and a Professor of Biostatistics at Harvard School of Public Health. Since 1999, Rafael Irizarry's work has focused on Genomics and Computational Biology problems. In particular, he has worked on the analysis and signal processing of microarray, next-generation sequencing, and genomic data. He is currently interested in leveraging his knowledge in translational work, e.g. developing diagnostic tools and discovering biomarkers.

Professor Irizarry also develops open source software implementing his statistical methodology. His software tools are widely used, and he is one of the leaders and founders of the Bioconductor Project, an open source and open development software project for the analysis of genomic data. Bioconductor provides one of the most widely used software tools for the analysis of microarray data.

DAY 5 – Friday 6 December

BIOC ASIA / PRECISION MEDICINE

08:45-09:00	**REGISTRATION**
09:00-09:40	How to Advance Science Using Bioconductor Professor Martin Morgan, Roswell Park Comprehensive Cancer Center
09:40-09:55	The RNAseq123 workflow package in Bioconductor Associate Professor Matthew Ritchie, Walter and Eliza Hall Institute of Medical Research
09:55-10:10	ClinSV: Detection of clinically relevant structural and copy number variation from whole genome sequencing Dr Andre Minoche, Kinghorn Centre for Clinical Genomics
10:10-10:15	Experiences of a First-Time Package Contributor Dr Stephen Pedersen, University of Adelaide
10:15-10:20	schex avoids overplotting for large single cell RNA-sequencing datasets Dr Saskia Freytag, Harry Perkins Institute of Medical Research
10:20-10:25	COmapR: Genetic length calculation from crossover events Ruqian Lyu
10:30-11:00	**MORNING TEA (catered)**
11:00-11:20	Single cell analysis with Mass Cytometry; technology introduction and opportunities in clinical studies Dr Helen McGuire, Charles Perkins Centre, The University of Sydney
11:20-12:00	On differential discovery in high-dimensional cytometry data Helena Crowell, University of Zurich
12:00-12:30	Ethics of precision medicine Dr Lisa Dive, The University of Sydney
12:30-13:30	**LUNCH**
13:30-15:30	WORKSHOPS jointly held with BioCAsia
Stream A:	Differential discovery in high-dimensional cytometry data Helena Crowell, University of Zurich
Stream B:	Fluent genomics: a plyranges and tximeta case-study Stuart Lee, Monash University, Walter and Eliza Hall Institute for Medical Research
Stream C:	Reproducible bioinformatics Dave Tang
15:30-15:50	**AFTERNOON TEA (catered)**
15:50-16:45	Defining immune signatures of therapeutic response with non-negative matrix factorization of bulk and single cell data Associate Professor Elana Fertig, Johns Hopkins University
16:45-17:00	**CONFERENCE CLOSE**

HOW TO ADVANCE SCIENCE USING BIOCONDUCTOR

Professor Martin Morgan, Roswell Park Comprehensive Cancer Center

Now is an incredibly exciting time for you to participate in bioinformatic research, with new methods for generating and analyzing large genomic data sets emerging almost daily. As bioinformaticians, we need a large, comprehensive, current, established, well supported, open source, community developed collection of software to address our leading-edge needs. This presentation describes the Bioconductor project, providing exactly these resources for the R bioinformatics community. We'll learn a little about the history, philosophy, and approach of the Bioconductor project. We then walk through essential steps in getting started with Bioconductor, tackling real analytic challenges, and contributing to the Bioconductor community. With luck, the presentation will inspire and empower you to tackle new and innovative challenges in your own bioinformatic research.

Martin earned his undergraduate and Master's degrees in Botany at the University of Toronto. Martin's PhD studies at the University of Chicago involved the evolutionary consequences of frequency-dependent selection, and of multilocus deleterious mutation. Martin is currently at the Roswell Park Comprehensive Cancer Center in Buffalo.

Martin leads the core team that maintains the Bioconductor project. He is the author of many Bioconductor packages and a renowned biostatistican.

THE RNASEQ123 WORKFLOW PACKAGE IN BIOCONDUCTOR

Associate Professor Matthew Ritchie, Walter and Eliza Hall Institute of Medical Research

The ability to easily and efficiently analyse RNA-sequencing data is a key strength of the Bioconductor project. In this presentation, I will introduce the RNAseq123 workflow package which demonstrates use of the popular edgeR package to import, organise, filter and normalise data, followed by the limma package with its voom method, linear modelling and empirical Bayes moderation to assess differential expression and perform gene set testing. This pipeline is further enhanced by the Glimma package which enables interactive exploration of the results so that individual samples and genes can be examined by the user. The complete analysis offered by these three packages highlights the ease with which researchers can turn the raw counts from an RNA-sequencing experiment into biological insights using Bioconductor.

https://bioconductor.org/packages/RNAseq123

CLINSV: DETECTION OF CLINICALLY RELEVANT STRUCTURAL AND COPY NUMBER VARIATION FROM WHOLE GENOME SEQUENCING

Dr Andre Minoche, Kinghorn Centre for Clinical Genomics

Clinical-grade detection and interpretation of structural variation (SV) including copy number variation (CNV) from whole genome sequencing (WGS) has the potential to revolutionize genetic testing by replacing micro-arrays. Current WGS based approaches however showed high error rates, poor reproducibility, and difficulties in annotating, visualizing, and prioritizing rare variants.

We present ClinSV, a platform addressing these challenges, by integrating read depth, split and spanning read evidence, with extensive quality attributes and, by providing a comprehensive variant visualization procedure. By analyzing WGS from 500 healthy individuals, ClinSV annotates calls with population allele frequencies, enabling filtration of on average 5,800 SVs down to 16 rare gene-affecting variants per germline sample. We benchmarked ClinSV against clinical microarrays and gold standard deletion CNV calls from NA12878, as well as other popular SV callers. ClinSV identified 100% of the pathogenic CNV from microarrays, including seven that were not detected using only a structural variant caller. Finally, we found that 13% of CNV had surrounding pairs of repeats, causing reduced SR and DP evidence, and highlighting the utility of integrating complementary CNV detection approaches.

EXPERIENCES OF A FIRST-TIME PACKAGE CONTRIBUTOR

Dr Stephen Pedersen, University of Adelaide

After being a long-term Bioconductor user and advanced R programmer, I thought I'd have package submission down. Turns out I was wrong and I learned many valuable and beneficial lessons.

SCHEX AVOIDS OVERPLOTTING FOR LARGE SINGLE CELL RNA-SEQUENCING DATASETS

Dr Saskia Freytag, UWA Centre for Medical Research

Visualizations, especially of dimension reductions, are the workhorse of all single cell RNA-sequencing (RNA-seq) data analyses. Currently employed visualization techniques struggle with the scale and sparsity of scRNA-seq data with sometimes disastrous consequences, such as masking or overrepresenting cells expressing a marker genes. To address this we developed the R package schex, which allows users to produce hexgonal binning representations for dimension reductions of scRNA-seq data stored in Seurat or SingleCellExperiment objects. Hexagonal binning representations, which summarize neighbouring points into hexagons, are a widely used technique when plotting extremely large datasets. We demonstrate that they can lead to more accurate plots that are at the same time faster and require less storage than traditional plots. This increase in speed of plotting also makes them especially suited towards interactive visualizations. Finally, we showcase some of schex's new capabilities that allow plotting of different modalities as well as interactions between modalities.

https://github.com/SaskiaFreytag/schex

COMAPR: GENETIC LENGTH CALCULATION FROM CROSSOVER EVENTS

Ruqian Lyu, St Vincent's Institute of Medical Research

Meiotic crossovers during spermatogenesis ensure genetic diversity in the haploid gametes and are tightly regulated. For example, FANCM has been shown to have a major role in limiting meiotic crossovers in Arabidopsis thaliana, fission yeast and Drosophila. Studies that investigate questions like this look at genetic lengths of organisms in different groups. Genetic lengths, measured in Morgan or centiMorgan, between two markers are derived from crossover rates via mapping functions with

different assumptions, i.e. Haldane or Kosambi. Organisms with more crossovers have larger total genetic lengths.

Recombination frequencies or crossover rates between two SNP markers are calculated by taking the ratio of counts of recombinants and counts of non-recombinants in a population. Genotyping of an array of SNP markers for a group samples allows the estimation of crossover rates and the calculation of a genetic map. The resolution of the map depends on the number and quality of genotyped markers.

Comapr is an R/Bioconductor package that converts genotyping test reports of genotypes for SNP markers in Excel spreadsheets to quantified genetic lengths. In future work, it will be reading other formats of input data such as VCF file.

In this package, we include functions for evaluating various quality metrics for SNP markers and filtering out low-quality markers. We also include functions for filtering samples when they have lots of missing data; appear to be duplicate etc. We also provide statistical testing functions for comparing genetic lengths between two populations or experimental groups and provide a significance value.

https://gitlab.svi.edu.au/biocellgen-private/comapr

SINGLE CELL ANALYSIS WITH MASS CYTOMETRY; TECHNOLOGY INTRODUCTION AND OPPORTUNITIES IN CLINICAL STUDIES

Dr Helen McGuire, Charles Perkins Centre, The University of Sydney

Mass cytometry, or Cytometry by Time-Of-Flight (CyTOF), is a powerful platform for high-dimensional single-cell analysis of the immune system. It enables the simultaneous measurement of over 40 markers on individual cells through the use of monoclonal antibodies conjugated to rare-earth heavy metal isotopes. Coupled with our already extensive immunological knowledge of canonical immune subsets and an ability to delve into and describe subtle populations, mass cytometry presents an opportunity to investigate cumulative subtle changes across many specific immune subsets in a range of clinical cohorts.

Based on our previous studies in several autoimmune states, which revealed remarkably stable changes in the size of multiple peripheral blood cell subsets, we conducted a study of cell subsets in melanoma and lung cancer patients before and after therapy with the checkpoint inhibitor, anti-PD-1. We used a data analysis approach originally developed to analyse gene expression signatures in highly multiparametric datasets to analyse the cell subset distribution within samples. We identified an immune signature in baseline blood samples that robustly identified patients who would subsequently make clinical responses to anti-PD-1 therapy. Such an approach is well suited to machine learning, which will be used in future application of the predictive signature in clinical settings.

Dr Helen McGuire is a Senior Research Officer at the Ramaciotti Facility for Human Systems Biology, Charles Perkins Centre, an initiative established in 2013 to support the development of mass cytometry and wider systems biology analysis across the University of Sydney campus and wider collaborative links. Her research focus and interest lies in the clinical application of immunological studies to a range of human diseases, including cardiovascular disease and cancer.

ON DIFFERENTIAL DISCOVERY IN HIGH-DIMENSIONAL CYTOMETRY DATA

Helena Crowell, University of Zurich

Mass cytometry (CyTOF) allows for examination of dozens of proteins at single-cell resolution. By employing heavy metal isotopes rather than fluorescent tags, thereby significantly reducing spectral overlap, CyTOF enables generation of high-throughput high-dimensional cytometry data.

Given the emergence of replicated multi-condition experiments, a primary task in the analysis of any type of single-cell data is to make sample-level inferences, in order to identify i) differentially abundant subpopulations; and, ii) changes in expression at the subpopulation-level, i.e., differential states (DS), across conditions. Preceding such analyses, key challenges lie in data preprocessing (e.g., to remove artefactual signal), clustering (to define subpopulations), and dimension reduction.

In this talk, I will present a suite of tools for differential discovery in CyTOF data, including 'CATALYST' for preprocessing and visualization, 'diffcyt' for differential testing, and a comprehensive analysis pipeline that leverages R/Bioconductor infrastructure. Secondly, I will cover benchmarks of key analysis steps, such as clustering and dimension reduction. Finally, I will touch on how we transferred our DS analysis framework to scRNA-seq, and developed a complex, flexible simulation framework for method comparison, with the 'muscat' package.

Helena earned her undergraduate degree at the Univeristy of Heidelberg in Biochemistry. She then went on to earn her Master's degree in Computational Biology & Bioinformatics at the ETH Zurich. She is currently a PhD candidate in Statistical Bioinformatics at the University of Zurich. Helena focuses on developing analysis frameworks for CyTOF data and differential discovery in scRNA-seq data. She is the author of a popular Bioconductor package providing tools for preprocessing and analysis of cytometry data.

ETHICS OF PRECISION MEDICINE

Dr Lisa Dive, The University of Sydney

With a background in analytic philosophy and professional experience in health policy, Lisa has a strong interest in applied ethics in the healthcare context. Her research explores the challenges that emerging medical technologies – such as genomics – pose for fundamental concepts in medical ethics. She currently has a particular focus on patient autonomy, examining how it comes under pressure in information-intensive areas of medicine, and the epistemic challenges that arise with increasing complexity of medical knowledge.

DIFFERENTIAL DISCOVERY IN HIGH-DIMENSIONAL CYTOMETRY DATA

Helena Crowell, University of Zurich

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.

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

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.

FLUENT GENOMICS: A PLYRANGES AND TXIMETA CASE-STUDY

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

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.

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

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.

REPRODUCIBLE BIOINFORMATICS

Dave Tang

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.

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

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.

DEFINING IMMUNE SIGNATURES OF THERAPEUTIC RESPONSE WITH NON-NEGATIVE MATRIX FACTORIZATION OF BULK AND SINGLE CELL DATA

Associate Professor Elana Fertig, Johns Hopkins University

In spite of advances to targeted therapies and immunotherapies, therapeutic resistance is still a critical challenge in cancer. The interactions between cancer and other cells in the tumor microenvironment drive therapeutic response and resistance. Time course genomics profiling and new single cell

technologies unprecedented measurements of the tumor microenvironment to elucidate the unknown interactions responsible for therapeutic response and resistance. Interpreting the molecular and cellular mechanisms of therapeutic response relies critically on the analysis methods to interpret these data. We demonstrate that the matrix factorization method CoGAPS uncovers regulatory mechanisms associated with therapeutic resistance in model systems. Furthermore, the patterns learned with CoGAPS can be related to patient response through transfer learning to deconvolve distance resistance mechanisms across data modalities and in patient samples.

Dr. Fertig runs a hybrid computational and experimental lab in the systems biology of cancer and therapeutic response. Her wet lab develops time course models of therapeutic resistance and performs single cell technology development. Her computational methods blend mathematical modeling and artificial intelligence to determine the biomarkers and molecular mechanisms of therapeutic resistance from multi-platform genomics data. These techniques have broad applicability beyond her resistance models, including notably to the analysis of clinical biospecimens, developmental biology, and neuroscience.

Dr. Fertig is an Associate Professor of Oncology and Assistant Director of the Research Program in Quantitative Sciences at Johns Hopkins University, with secondary appointments in Biomedical Engineering and Applied Mathematics and Statistics, affiliations in the Institute of Computational Medicine, Center for Computational Genomics, Machine Learning, Mathematical Institute for Data Science, and the Center for Computational Biology. Prior to entering the field of computational cancer biology, Dr Fertig was a NASA research fellow in numerical weather prediction. Dr. Fertig's research is featured in numerous peer reviewed publications, R/Bioconductor packages, and independent research funding. She led the team that won the HPN-DREAM8 algorithm to predict phospho-proteomic trajectories from therapeutic response in cancer cells.