AB^{3}ACBS 2016

A festival of bioinformatics

COMBINE

COMBINE16 Symposium
AB^{3}ACBS 2016 Conference
DELEGATE HANDBOOK

31 October – 2 November 2016

Queensland University of Technology (QUT)
Gardens Point campus

www.abacbs.org/conference
combine.org.au/symposium-2016/

Proudly supported by Gold Sponsors:

illumina®

ADVANCE QUEENSLAND

AB^{3}ACBS 2016 is hosted by:

QUT Queensland University of Technology

Welcome

The organising committee welcomes you to Brisbane for the AB^{3}ACBS 2016 meeting, which this year integrates several events for bioinformaticians and computational biologists:

- The COMBINE symposium, the official student sub-committee of ABACBS
- The ABACBS Annual Conference
- B^3, the Big Biology and Bioinformatics symposium
- GOBLET, the Global Organisation of Bioinformatics Learning, Education and Training
- Bioconductor through the BioCAsia meeting and training workshop

The weeklong AB^{3}ACBS 2016 meeting will be scientifically engaging, and will include a range of presentations from scientists across all stages of career. The COMBINE symposium and ABACBS conference will conclude with prizes being awarded to the best student posters and oral presentations. The training workshops will be rewarding and will enlighten attendees on best practices and current approaches for analysis and we are excited that GOBLET is part of the event. The social program includes the COMBINE trivia night and the ABACBS dinner, which will be held by the Brisbane River.
We are sure you will find the meeting rewarding and stimulating and look forward to welcoming you to Brisbane and QUT.

Sincerely,

Dr Tony Parker – Co Convenor
Queensland University of Technology
Email: a.parker@qut.edu.au
Phone: +61 7 3138 6187

Dr Nic Waddell – Co Convenor
QIMR Berghofer Medical Research Institute
Email: nic.waddell@qimrberghofer.edu.au
Phone: +61 7 3845 3538

Professor David Lovell – Chair Organising Committee
Queensland University of Technology
Email: david.lovell@qut.edu.au
Phone: +61 7 3138 1678

COMBINE16 Organising committee
Dr Paula Martinez
Ms Leah Roberts
Mr Zac Gerring
Ms Shila Ghazanfar
Ms Crystal Concepcion
Mr Joel Boyd
Mr Klay Saunders
Ms Sarah Williams
Mr Glen Van der Bergen
Ms Nesli Avgan

AB^{3}ACBS Organising committee
Dr Brett Williams
Dr Ana Pavasovic
Dr Peter Prentis
Ms Leah Roberts
Dr Paula Martinez
Mr Zac Gerring
Dr Annette McGrath
A/Prof Jim Hogan
Dr Mirana Ramialison
Mr John Pearson
Professor Mark Ragan

QUICK REFERENCE PROGRAM

Monday, 31 October 2016

COMBINE16 STUDENT SYMPOSIUM

Time	Activity
8.30-9.30am	COMBINE registration and poster hanging
9.00-9.15am	Welcome and opening remarks
9:15-9:55am	Talk Session 1: • **Soroor Hediyeh-Zadeh** Computational workflows for research students: towards a reproducible research • **Nesli Avgan** Applying machine learning to GWAS analysis of major memory traits
9:55-10:25am	Poster Lightning Talks 1
10:25-11:00am	Morning tea and poster viewing
11:00-12:00pm	Talk Session 2: • **Momeneh Foroutan** A transcriptional signature for TGFβ-induced epithelial-mesenchymal transition in cancer • **Luke Zappia** Simplifying simulation of single-cell RNA sequencing • **Ramyar Molania** Accurate and robust normalization of Nanostring nCounter gene expression data
12:00-1:15pm	Lunch and poster viewing
1:15-1:30pm	Group photo
1:30-2:30pm	Talk Session 3: • **Andrew Pattison** Investigating the 3' untranslated region of mRNA in order to understand the drivers of metastasis in primary triple negative breast tumours • **Areej Alsheikh-Hussain** Insertion sequence elements are drivers of diversification in the broad host range aquatic pathogen Streptococcus iniae • **Son Hoang Nguyen** Scaffolding and completing genome assemblies in real-time with nanopore sequencing
2:30-3:00pm	Poster Lightning Talks 2
3:00-3:30pm	Afternoon tea at Botanic Bar
3:30-5:00pm	Panel discussion • **David Lovell, Katherine Pillman, Lynn Fink, Nic Waddell and CX Chan**
5:00-5:15pm	Talk and poster prizes, Symposium close
5:15-6:30pm	Free time
6:30pm- late	COMBINE Trivia night and social event - Botanic Bar

AB^{3}ACBS CONFERENCE, DAY 1

Tuesday, 1 November

Time	Activity
8:00-9:00am	AB^{3}ACBS registration and poster hanging
9:00-9:15am	Housekeeping; Welcome; and Opening by **Queensland Chief Scientist – Dr Geoff Garrett**
9:15-10:30am	**SESSION 1:** Microbes
	• **Nouri Ben Zakour** Evolutionary epidemiology of successful bacterial pathogens • **Andrew Buultjens** Exploiting extremes of Legionella pneumophila genomic diversity for accurate source attribution • **Darryl Reeves** Probabilistic Inference and Shared Parameter Learning for Metagenomic Sequence Analysis • **Patrick Laffy** HoloVir: Taxonomic and functional analysis of viral metagenomic communities.
10:30-11:00am	Morning tea
11:00-12:30pm	**SESSION 2:** Cancer
	• **Andreas Schreiber** Taking the confusion out of fusions: Structural mutation detection in cancer research and diagnostics • **Shila Ghazanfar** Exploring Pan-Cancer Network Relationships Between Somatic Changes and Expression Profiles with PACMEN • **Luis Lara-Gonzalez** Consequences of Drug Dose Modulation on Clonal Dynamics • **Rebecca C. Poulos** Functional mutations form at CTCF/cohesin binding sites in melanoma due to uneven nucleotide excision repair across the motif • **Christoffer Flensburg** Tracking clonal evolution in cancer from multiple samples
12:30-1:30pm	Lunch
1:30-3:00pm	**SESSION 3:** Statistics
	• **Matt Ritchie** Tools for comparing and combining RNA-seq results • **Paul Lin** CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-Seq data • **Belinda Phipson** Gene length bias in single cell RNA-seq data • **Joseph Cursons** Computational methods to examine micro-RNA targeting of interacting protein networks • **Göknur Giner** FRY: A Fast Approximation to ROAST Gene Set Test with Mean Aggregated Set Statistics
3:00-3:30pm	Afternoon tea
3:30-5:00pm	**SESSION 4:** Plants and Animals
	• **Kate Hertweck** Plant systematics to cancer biology: genome-wide patterns and organismal evolution • **Jimmy Breen** Whole methylome analysis of grapevine reveals tissue specific DNA methylation variation • **Andrew Lonsdale** Mighty Morphin FASTA Files • **Nadia Davidson** SuperTranscript: a compact reference for the transcriptome • **Stephen Fletcher** Small Complementary RnA Mapper (SCRAM): a tool for studying small interfering RNA biogenesis in plants
5:00pm	**Poster Session**
6:30pm	**Conference Dinner**

AB^{3}ACBS CONFERENCE, DAY 2

Wednesday, 2 November

Time	Activity
9:00-10:30am	**SESSION 5:** Evolution
	• **Simon Ho** Phylogenomic analysis and molecular evolutionary clocks • **Åsa Pérez-Bercoff** Investigating the evolution of new biochemical pathways in baker's yeast Saccharomyces cerevisiae • **Miles Benton** Identification of allelic-specific methylation profiles across generations in the Norfolk Island genetic isolate • **Kevin Murray** Estimating genetic similarity with the k-mer Weighted Inner Product (kWIP) • **Martin Smith** De novo characterisation of RNA structure motifs from ENCODE RIPseq data
10:30-11:00am	Morning tea
11:00-12:30pm	**SESSION 6:** ‘Omics
	• **Ute Roessner** Metabolomics - an important piece in the 'omics puzzle • **Pip Griffin** EMBL Australia Bioinformatics Resource (EMBL-ABR) • **Jovana Maksimovic** Through the looking glass: using Monocle to visualise methylation array data • **Ruth Fuhrman-Luck** Big data from a little proteolysis: Combining multiple omics platforms to identify novel functions of a protease associated with prostate cancer • **Lawrence Buckingham** Geometric map-reduce algorithms for alignment-free sequence comparison
12:30-1:00pm	Lunch
1:00-2:00pm	ABACBS Annual General Meeting
2:00-3:40pm	**SESSION 7:** Disease
	• COMBINE Best Student Talk • **Daniel Cameron** GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly • **Harriet Dashnow** Detecting pathogenic STR expansions in next-gen sequencing data • **Rod Lea** BootNet: a bootstrapping application for GLMnet modelling to identify robust classifiers in genomics data • **Alexis Lucattini** Assessing the Practicality of Oxford Nanopore Sequencing in Clinical Diagnositcs
3:40-5:00pm	**SESSION 8**: Systems
	• **Anna Trigos** How do vulnerabilities left during the evolution of cellular networks set the stage for cancer? • **Stuart Archer** TCP-seq, a novel technique for investigating mechanisms and regulation of eukaryotic translation initiation • **Teresa Attwood** The Road to Utopia: challenges in linking literature & research data
5:00pm	**Close**

GOBLET BEST PRACTICES IN BIOINFORMATICS TRAINING, Day 1

Thursday, 3 November

Time	Activity
9:00-9:15am	**Welcome and housekeeping – Annette McGrath (CSIRO)**
9:15-10:00am	**Clinical Bioinformatics**
	• Experiences in teaching bioinformatics for clinical audiences **Gabriella Rustici**, *University of Cambridge* • Australian experiences in teaching bioinformatics for the clinic **Bronwyn Terrill**, *Garvan Institute of Medical Research*
10:30-11:00am	Morning tea
11:00-12:30pm	**ISCB Core Competencies**
	• Taking ISCB competencies forward to using the competencies for course/program/curriculum design **Bruno Gaeta**, *UNSW*
12:30-1:30pm	Lunch
1:30-2:30pm	**Teaching programming**
	• An R course developed by the Monash Bioinformatics Platform **Paul Harrison**, *Monash Bioinformatics Platform* • Software Carpentry **Belinda Weaver**, *QCIF/Software Carpentry* • Software Carpentry and Data Carpentry in Bioinformatics Training **Harriet Dashnow**, *Murdoch Children's Research Institute* • Enabling community training and support: successes and challenges from the Bioconductor project **Martin Morgan**, *Bioconductor*
2:30-300pm	**Q&A/Panel Discussion**
3:00-3:30pm	Afternoon tea
3:30-4:30pm	**Use of clouds/VMs in bioinformatics training**
	• Introduction to clouds, terminology & resources **Steve Androulakis**, *Monash Bioinformatics Platform* • The Genomics Virtual Laboratory as a training platform **Igor Makunin**, *UQ RCC* • BPA/CSIRO cloud training platform **Jerico Revote**, *Monash University* • Outcomes of the 3 day workshop In Europe on the use of cloud/VMs for training **Eija Korpelainen**, *CSC-IT Center for Science, Finland*
4:30-5:00pm	**Q&A/Panel Discussion**
5:00-5:30pm	How can the ABACBS Education committee and GOBLET support our training community?
5:30pm	**Close**

INTRODUCTION TO BIOCONDUCTOR/SHINY WORKSHOP

Thursday, 3 November Q228

Time	Activity
9:00-10:30am	**Intro to Bioconductor Session 1** **Martin Morgan**, *Roswell Park Cancer Institute*
10:30-11:00am	Morning tea
11:00-12:00am	**Bioconductor Session 2** **Martin Morgan**, *Roswell Park Cancer Institute*
12:00-1:00pm	Lunch
1:00-2:00pm	**Bioconductor Session 3** **Martin Morgan**, *Roswell Park Cancer Institute*
2:00-3:00pm	**Shiny Session 1** **Paul Harrison***, Monash Bioinformatics Platform*
3:00-3:30pm	Afternoon tea
3:30-5:00pm	**Shiny Session 2** **Paul Harrison***, Monash Bioinformatics Platform*
5:00pm	**Close**

GOBLET BEST PRACTICES IN BIOINFORMATICS TRAINING, Day 2

GOBLET/ELIXIR TtT session

Friday, 4 November

Time	Activity
9:00-9:10am	**Welcome and Introduction – Sarah Morgan (EBI)**
9:10-10:30am	**Train-the-trainer (TtT) session**
	• Good vs bad training the trainers **Annette McGrath**, *CSIRO*
10:30-11:00am	Morning tea
11:00-12:30pm	• How do people learn? **Sonika Tyagi**, *AGRF/EMBL-ABR*
12:30-1:30pm	Lunch
1:30-3:00pm	• Session and course design **Ann-Marie Patch**, *QIMR Berghofer*
3:00-3:30pm	Afternoon tea
3:30-5:00pm	• Gathering feedback **Sean McWilliam**, *CSIRO*
5:00pm	**Close**

SECOND ASIA-PACIFIC BIOCONDUCTOR MEETING

Friday, 4 November, Q228

Time	Activity
9:00-9:55am	Project Updates (**Prof Martin Morgan**)
9:55-10:30am	**Lightning talks** **Soroor Hediyeh Zadeh** 'OPPAR: Outlier Profile and Pathway Analysis' **Ted Wong** 'Anaquin - quantitative controls with spike-ins' **Jovana Maksimovic** 'A Bioconductor Workflow for Methylation Array Analysis' **Jason Ross** 'The Development of the R/Bioconductor build of Harman' **Dario Strbenac** 'ClassifyR: Classification Convenience in R' **Goknur Giner** 'Introducing R-LadiesAU'
10:30-11:00am	Morning Tea
11:00-12:30pm	Research talks (4 x 20 minutes each) Session Chair: Jovana Maksimovic **Belinda Phipson** 'Analysing DNA methylation array data with missMethyl: going beyond differential methylation' **Tim Peters** 'DMRcate, a complete DMR-calling Bioconductor package' **Andy Chen** 'Analyzing Bisulfite sequencing methylation data using edgeR' **Hien To** 'rnaCleanR: A tool for quantifying and removing DNA contamination from strand-specific RNA-seq'
12:30-1:20pm	Lunch
3:00-3:30pm	Research talks (5 x 20 minutes each) Session Chair: Stephen Pederson **Charity Law** 'Glimma, getting greater graphics for your genes' **Davis McCarthy** 'scater: pre-processing, quality control, normalisation and visualisation for single-cell RNA-seq data' **Goknur Giner** 'FRY: A Fast Approximation to ROAST Gene Set Test with Mean Aggregated Set Statistics' **Monther Alhamdoosh** 'Combining multiple tools outperforms individual methods in gene set enrichment analyses' **Cynthia Liu** 'KRLMM: a SNP genotyping method for both common and low-frequency variants in any organism'
3:00-3:30pm	Afternoon Tea
3:30-4:10pm	Keynote: Keynote: **Gordon Smyth** 'A short history of limma and edgeR'
4:10-4:30pm	Discussion and closing remarks

MIXOMICS WORKSHOP

Multivariate data analysis methods for biological data

Friday, 4 November, P638

Time	Activity
9:00am	**Start**
10:30-11:00am	**Coffee break**
12:30-1:30pm	**Lunch**
3:00-3:30pm	**Coffee break**
5:00pm	**End**

Overview:

Multivariate dimension reduction approaches are useful exploratory tools to get a first understanding of large and complex data sets. These approaches are extremely efficient to compute and highly flexible as they can answer a variety of biological integrative questions. Our latest developments in that exciting area of research include feature selection and statistical integration of several `omics data sets.

The workshop will introduce the fundamental concepts of multivariate dimension reduction methodologies for data exploration, identification of biomarkers and integration of large data sets; especially in the context of systems biology, or in research areas where statistical data integration is required. Each methodology that will be introduced will be applied on biological 'omics studies including transcriptomics, metabolomics and proteomics data sets using the R package mixOmics (http://mixomics.org/).

Presenter:

Dr Kim-Anh Lê Cao (The University of Queensland Diamantina Institute, Brisbane Australia) is an expert in multivariate statistical methods and novel developments. Since 2009, her team has been working on developing mixOmics dedicated to the integrative analysis of `omics' data, to help researchers mine and make sense of biological big data (http://www.mixOmics.org).

Tutors:

Dr Florian Rohart (University of Queensland Diamantina Institute, Brisbane Australia) is a core developer for mixOmics and develops cutting-edge multivariate methods for horizontal and vertical integration.

Mr **Nicholas Matigian** (Biostatistics facility, The University of Queensland Diamantina Institute) is specialized in high-throughput data analysis and provide assistance and support to UQDI researchers.

Monday 31 October

COMBINE Student Symposium

9.00am – 5.15pm, Level 5, P Block

Time	Activity
8.30-9.30am	COMBINE registration and poster hanging
9.00-9.15am	Welcome and opening remarks
9:15-9:55am	Talk Session 1: • **Soroor Hediyeh-Zadeh** Computational workflows for research students: towards a reproducible research • **Nesli Avgan** Applying machine learning to GWAS analysis of major memory traits
9:55-10:25am	Poster Lightning Talks 1
10:25-11:00am	Morning tea and poster viewing
11:00-12:00pm	Talk Session 2: • **Momeneh Foroutan** A transcriptional signature for TGFβ-induced epithelial-mesenchymal transition in cancer • **Luke Zappia** Simplifying simulation of single-cell RNA sequencing • **Ramyar Molania** Accurate and robust normalization of Nanostring nCounter gene expression data
12:00-1:15pm	Lunch and poster viewing
1:15-1:30pm	Group photo
1:30-2:30pm	Talk Session 3: • **Andrew Pattison** Investigating the 3' untranslated region of mRNA in order to understand the drivers of metastasis in primary triple negative breast tumours • **Areej Alsheikh-Hussain** Insertion sequence elements are drivers of diversification in the broad host range aquatic pathogen Streptococcus iniae • **Son Hoang Nguyen** Scaffolding and completing genome assemblies in real-time with nanopore sequencing
2:30-3:00pm	Poster Lightning Talks 2
3:00-3:30pm	Afternoon tea at Botanic Bar
3:30-5:00pm	Panel discussion • **David Lovell, Katherine Pillman, Lynn Fink, Nic Waddell and CX Chan**
5:00-5:15pm	Talk and poster prizes, Symposium close
5:15-6:30pm	Free time
6:30pm- late	COMBINE Trivia night and social event - Botanic Bar

ORAL PRESENTATION ABSTRACTS

Computational workflows for research students: towards a reproducible research

Soroor Hediyeh-Zadeh and Melissa J Davis

The Walter and Eliza Hall Institute of Medical Research

The majority of new bioinformatics research students are often not acquainted with the common computational workflows and resources. Students may spend several months writing scripts to perform routine analysis, for which a comprehensive set of tools might have already been developed.

Bioconductor is an open source Bioinformatics software (in R language) repository, which provides tools and resources for almost any type of Bioinformatics data analysis, from sequence analysis and differential gene expression to proteomics and metabolomics data analysis. As a very simple example, Bioconductor provides data structures to store gene expression data, the associated gene annotation and the sample information all in one R object (e.g. ExpressionSet object for microarray data, and summarizedExperiment object for RNA-Seq/count data). This ensures consistency when subsetting is done on one table. In addition to tools for reading and pre-processing different data formats (e.g. CELL, SAM/BAM, GTF/GFF, FASTA/FASTQ etc.) many of the command-line tools are now available through Bioconductor, and can be accessed directly from R. It is also possible to retrieve the genomic sequence/location and annotation data for a vast number of species and genome builds. Other public datasets provided by UCSC, NCBI, Refseq etc. can also be accessed directly via Bioconductor packages. In addition, GEO datasets can be downloaded and accessed directly from an R session. All these different tools minimize manual file handling and ensure a reproducible research.

In this talk/poster, the core Bioconductor technologies are introduced. The term Reproducible Research is defined. I, then, introduce Docker and explain how it can be deployed to build a digital archive that reproduces the research results. Finally, the advantages and disadvantages of using Docker over other tools for research reproducibility are discussed.

Applying machine learning to GWAS analysis of major memory traits

Nesli Avgan (a), Rodney A Le (a), Miles Benton (a), Heidi G Sutherland (a), Lauren G Spriggens (b) David HK Shum (b), Larisa M Haupt (a), Lyn R Griffiths (a)

(a) Genomics Research Centre, Chronic Disease and Ageing, Institute of Health and Biomedical Innovation, School of Biomedical Sciences, Queensland University of Technology, Brisbane, Australia (b) Behavioral Basis of Health Program, Menzies Health Institute Queensland, Griffith University, Gold Coast, Australia

Genome wide association studies (GWAS) is a well-established design for identifying genetic variants that are associated with complex common disease traits. However, there are statistical limitations in GWAS studies, such as multiple testing burden and reduction of power, detecting the effects of SNPs separately and small effect size of the detected SNPs and integrating gene-gene and gene-environmental interactions. In an effort to overcome these limitations we are applying a machine learning approach called GLMNet. Briefly, GLMNet offers extremely efficient procedures for fitting the entire lasso or elastic-net regularization path for regression models. We will apply GLMNet to SNP data for a GWAS study of 619 healthy individuals measured for a battery of tests for evaluating twenty different memory performance phenotypes including visual, episodic and prospective memory. We will compare the GLMNet results to our results of a conventional analysis, performed using PLINK. These results showed highly significant SNP associations for several memory traits but failed to explain a high proportion of genetic heritability of these traits. Here, our goal is to use GLMNet to identify SNP signatures that are highly predictive of these complex memory traits and in turn improve the biological understanding of human memory.

A transcriptional signature for TGFβ-induced epithelial-mesenchymal transition in cancer

Momeneh Foroutan (a,b), Joseph Cursons (b,c,d), Soroor Hediyeh-Zadeh (b), Erik W. Thompson (a,e,f), Melissa J. Davis (b,g)

(a) The University of Melbourne Department of Surgery, St. Vincent's Hospital, Parkville, VIC 3010, AUSTRALIA; (b) Division of Bioinformatics, Walter and Eliza Hall Institute of Medical Research, Parkville, VIC 3010, AUSTRALIA; (c) Systems Biology Laboratory, Melbourne School of Engineering, The University of Melbourne, Parkville, VIC 3010, AUSTRALIA. (d) ARC Centre of Excellence in Convergent Bio-Nano Science and Technology, Melbourne School of Engineering, The University of Melbourne, Parkville, VIC 3010, AUSTRALIA. (e) Institute of Health and Biomedical Innovation and School of Biomedical Sciences, Queensland University of Technology, Kelvin Grove, QLD 4059, AUSTRALIA (f) Translational Research Institute, Wooloongabba, QLD, 4102, AUSTRALIA (g) Department of Biochemistry and Molecular Biology, Faculty of Medicine, Dentistry and Health, University of Melbourne, Parkville, VIC 3010, AUSTRALIA.

Epithelial-mesenchymal transition (EMT) is a developmental program subverted by cancer cells as they progress through the metastatic cascade. EMT is associated with an aggressive, migratory cell phenotype and it can contribute to clinical chemotherapeutic resistance and poor treatment outcomes. EMT can be induced by different stimuli, of which transforming growth factor β (TGFβ) is one of the best studied. We have used meta-analysis methods to identify a robust transcriptomic signature of TGFβ-induced EMT, and then used this signature to distinguish cancer cell lines and patient samples with evidence of TGFβ-induced EMT. We examine the signature across multiple breast cancer and pan-cancer datasets, demonstrating that: our results are reproducible on independent data; cell-line and patient samples show consistent, cancer type-specific levels of TGFβ-EMT activity, and; our TGFβ-induced EMT signature is influenced by the accumulation of genetic mutations across the TGFβ signalling pathway. Finally, we apply our signature to stratify patients and show differences in survival outcome, and identify cell lines with resistance to common cancer drugs.

Simplifying simulation of single-cell RNA sequencing

Luke Zappia (a,b), Belinda Phipson (a) and Alicia Oshlack (a,b)

(a) Bioinformatics, Murdoch Childrens Research Institute; (b) School of Biosciences, The University of Melbourne

Single-cell RNA sequencing (scRNA-seq) is rapidly becoming a tool of choice for biologists who wish to investigate gene expression, particularly in areas such as development and differentiation. In contrast to traditional bulk RNA-seq experiments, which measure expression averaged across millions of cells, single-cell experiments can be used to observe how genes are expressed in individual cells. Along with the dramatic increase in resolution provided by scRNA-seq comes an array of bioinformatics challenges. Single-cell data is relatively sparse (for both biological and technical reasons), quality control is difficult and it is unclear how to replicate measurements. The focus of analysis is also different, with more emphasis on clustering cells to identify cell types or ordering of cells to understand dynamic processes than traditional tasks such as differential expression testing. Any new bioinformatics method for scRNA-seq analysis should demonstrate two things: 1) it can do what it claims and 2) it helps to produce biological insight. The first is hard to prove on real data where there is often no known truth. Because of this, bioinformaticians turn to simulations. Unfortunately current scRNA-seq simulations are frequently poorly documented, not reproducible and do not demonstrate similarity to real data or experimental designs. Here we discuss some of the problems with simulating scRNA-seq data and provide a simulation framework that addresses these concerns.

Accurate and robust normalization of Nanostring nCounter gene expression data

Ramyar Molania(a,b), Terence P Speed(c,d), Alexander Dobrovic(e,f)

(a) Department of Medicine, Austin Health, University of Melbourne; (b) Translational Genomics and Epigenomics Laboratory, Olivia Newton-John Cancer Research Institute; (c) Bioinformatics Division, Walter and Eliza Hall Institute of Medical Research; (d) Department of Mathematics and Statistics, University of Melbourne; (e) Department of Pathology, University of Melbourne; (f) School of Cancer Medicine, La Trobe University

Nanostring nCounter is being increasingly used for research and clinical studies due to its capability to directly measure gene expression for a large panel of genes from RNA, even when the RNA is substantially degraded. Most suggested normalization methods including the Nanostring nSolver software are based on using small panels of user-chosen housekeeping genes and Nanostring spike-in controls. These are insufficient to remove batch and other technical effects, particularly when the data comes from large or complex experiments.

We used a new normalization method, Removing Unwanted Variation III (RUVIII), which is a variant on the RUV method using replicate samples as published by Jacob et al, Biostatistics 2015. We compared the performance of RUVIII with other normalization methods using several evaluation criteria including similarity of replicate samples, box and Relative Log Expression (RLE) plots of un-normalized and normalized expression values, differential expression analyses and clustering methods including tSNE and PCA.

We applied the RUVIII method to two different in-house data sets, as well as 8 different studies with nearly 8000 samples from publicly available repositories. The results show the performance of RUVIII is markedly better than existing methods. In particular, the performance of RUVIII in dealing with complex experiments with substantial batch effects is significantly superior compared to other methods. RUVIII not only removes batch effects in the data sets, but also reveals otherwise masked biology in the experiments.

In conclusion, removing unwanted variation plays an essential role in obtaining precise results for gene expression data. Using panels of housekeeping genes to normalize Nanostring gene expression data may lead to unsatisfactory normalization and accordingly, misleading results. RUVIII gives superior performance as judged both metrics and by concordance with known biological findings.

Investigating the 3’ untranslated region of mRNA in order to understand the drivers of metastasis in primary triple negative breast tumours

Andrew Pattison (a), Cameron Johnstone (a,b), Paul Harrison (a,c), Robin Anderson (b), David Powell (c) and Traude Beilharz (a)

(a) Department of Biochemistry and Molecular Biology, Monash University, Melbourne, Australia; (b) Metastasis Research Laboratory, Olivia Newton-John Cancer Research Institute, Melbourne, Australia; (c) Monash Bioinformatics Platform, Monash University, Melbourne, Australia

While progress has been made in the detection of primary breast tumours, determining how likely a tumour will be to metastasise is still very difficult. Understanding this “metastatic potential” is most important in the so called “triple negative breast cancers” (TNBCs) which lack the classical markers that are commonly targeted in treatment. Chemotherapy is usually given for triple negative tumours, often unnecessarily. Better markers of tumour metastatic potential are clearly required.

Alternative polyadenylation (APA) is the process whereby the poly (A) tail is added to the 3’ untranslated region (3’ UTR) of mRNA at one of multiple possible sites, changing 3’ UTR length and potentially the regulatory elements that bind to it. APA has been shown to be indicative of tumour state, but is often overlooked when conducting RNA-seq analysis. We are developing a method to cheaply and effectively quantify the expression state of a primary breast tumour based off Poly (A) Test sequencing (PAT-seq) data, which sequences 3' UTRs in a genome wide fashion. PAT-Seq is also capable of measuring differential poly (A) tail length which may also play a role in metastasising tumour cells.

We hypothesise that the metastatic potential of a primary tumour can be calculated from changes in gene expression, APA site usage and the length of the poly (A) tail. We are testing this hypothesis in an increasingly metastatic cell line model both in vitro and in vivo. The model was generated from MDA-MB-231 TNBC cells and differences should be associated with metastatic potential. We have discovered some interesting 3' UTR associated changes in immune signalling, calcium ion balance and glutamine metabolism genes.

In order to effectively interpret and visualise PAT-seq data we make use of “Tail-tools”, a custom bioinformatics pipeline that employs modified versions of Limma and Degust. Recently, we have also begun to use Shiny to visualise PAT-seq data in new ways, often on a per gene basis.

Insertion sequence elements are drivers of diversification in the broad host range aquatic pathogen Streptococcus iniae

Areej Alsheikh-Hussain (c), Nouri L. Ben Zakour (a), Andrew C. Barnes (b) and Scott A. Beatson (c)

(a) The Westmead Institute for Medical Research, Sydney, Australia
(b) School of Biological Sciences and Centre for Marine Science, University of Queensland, Brisbane, Australia
(c) School of Chemistry and Molecular Biosciences, Australian Centre for Ecogenomics, University of Queensland, Brisbane, Australia

Fish mortality caused by Streptococcus iniae is a major economic problem in fresh and seawater finfish aquaculture in warm and temperate regions globally, including Australia and Southeast Asia. There is also risk of occasional zoonotic infection by S. iniae through handling of contaminated fish. Vaccination of fish against S. iniae sometimes fails due to the rapid mutation of genes in the capsular biosynthesis operon of S. iniae, which leads to novel serotypes and consequent re-infection of immunized stock. Currently, genomic analysis of S. iniae is scant, with only 4 complete genomes sequenced to date, although high strain diversity has been described using pulsed-field gel electrophoresis of genomic DNA. The role of mobile genetic elements, including insertion sequences (IS), in this diversification of S. iniae lineages or as mechanisms of adaptation has not yet been investigated. The aim of the present study is to investigate the evolutionary history of S. iniae and its adaptation mechanisms that enable such widely disseminated infection in different fish species, environments, and humans. To achieve this we sequenced a collection of 113 S. iniae isolates from different hosts of worldwide origin. A phylogenetic tree constructed by maximum likelihood analysis of alignment of non-recombinant core-genome SNPs from 113 strains in the collection, along with the four complete genomes available on Genbank revealed separate clustering of human and fish isolates, as well as phylogeographic grouping even within Australia, with strains from Queensland clustering separately to those from the Northern Territory. The distribution of IS elements was found to be clade-specific, where the CRISPR/Cas system was targeted by different IS types causing interruption of cas9, cas1, and csn2, some of which are homoplasies. This broad scale assessment of diversity and diversification at the genomic level forms the basis for identification of stable, conserved potential antigens for future development of improved vaccination against S. iniae and this work is now underway.

Scaffolding and completing genome assemblies in real-time with nanopore sequencing

Son Hoang Nguyen, Minh Duc Cao, Devika Ganesamoorthy, Tania Da Silva Duarte, Alysha G. Elliott, Matthew A. Cooper, Lachlan J.M. Coin

Institute for Molecular Bioscience, the University of Queensland

Genome assemblies using short read sequencing technology are often fragmented into many contigs because of the abundance of repetitive sequences. Long read sequencing technologies can generate reads spanning almost repeats, offering the opportunity to complete these genome assemblies. However, existing assembly methods require substantial amounts of sequence data and computational resources to overcome the high error rates in the long read data. Furthermore, these approaches only assemble genomes after sequencing has completed, which could result in either the generation of excessive sequence data at greater cost or a low-quality assembly due to data shortage.

Here we present npScarf, the first computational method utilizing real-time nanopore sequencing to scaffold and complete short read assemblies while the long read sequence data are being generated. The method reports the progress of completing the assembly in real-time so users can terminate sequencing once an assembly of sufficient quality and completeness is obtained.

We used npScarf to assemble four bacterial genomes and one eukaryotic genome, and showed that it was able to construct more complete and more accurate assemblies, while at the same time requiring less sequencing data and computational resources than existing pipelines. We ran npScarf to scaffold an MRSA sample directly from a MinION sequencing run to produce a near complete genome in 2 hours of sequencing. We then used the same MinION flowcell to complete a multi-drug resistance Klebsiella pneumoniae strain in 3 hours. We demonstrate that the method can facilitate real-time analyses of positional information such as identification of bacterial genes encoded in plasmids and pathogenicity islands. We also develop a visualization of the scaffolding on a web browser to report the progress of the analysis in real-time.

COMBINE LIGHTNING TALKS

	Presentation
1	Identification of the dominant endogenous factors regulating inflammation and regeneration in skeletal muscle following physical trauma **Lian Liu** *Queensland University of Technology*
2	Chromosome end extension revealed by analysis of completed chromosome end sequences **Haojing Shao** *The University of Queensland*
3	Identification of phylogenetically useful loci **Bokyung Choi** *The Australian National University*
4	De novo assembly of Sugarcane leaf RNAseq data: A cluster and merge approach **Kate Wathen-Dunn** *The University of Queensland*
5	Dimensionality reduction by t-SNE uncovers the connections that shape the landscape of the transcriptome **Michael See** *Monash University*
6	Quantification of paediatric burn blister fluid proteome using SWATH MS to assist better clinical diagnosis **Tuo Zang** *The Queensland University of Technology*
7	In silico analysis of immunomodulatory vaccine candidate proteins SpyCEP and EndoS in Streptococcus pyogenes **Lochlan Fennell** *QIMR Berghofer Medical Research Institute*
8	Finding optimal coverage **Anna Quaglieri** *The Walter and Eliza Hall Institute of Medical Research*
9	Cancer progression and hypoxia development of a pan cancer hypoxic-signature **Kristy Horan** *The Walter and Eliza Hall Institute of Medical Research*
10	Biomarkers of Anterior Cruciate Ligament Injury and Recovery **Yee Chng** *Queensland University of Technology*
11	eQTL analysis of quantitative endophenotypes for ocular health in the Norfolk Island isolate **Pik Fang Kho** *Queensland University of Technology*
12	Combining high throughput sequencing data to improve identification of transcription factor binding **Alex Essebier** *The University of Queensland*
13	Colorectal Cancer Atlas and FunRich: Discovery tools for integrated ‘omics’ data analysis **Naveen Chilamkurti** *La Trobe University*

PANEL DISCUSSION PROFILES

Professor David Lovell

Professor David Lovell is the Head of QUT's School of Electrical Engineering and Computer Science (EECS) from December 2014. David graduated with a BEng Computer Systems, Hons I from the University of Queensland 1989 and received his PhD there in 1994 for research into artificial neural network methods for handwritten character recognition. He completed postdoctoral research at Cambridge University before joining CSIRO Mathematical and Information Sciences (CMIS) in 1998. At CSIRO, David was involved in a wide range of research and consulting in the analysis of large and complex datasets, worked as Executive Officer to the CEO from 2001-02, and was a member of CSIRO's Corporate IT Management team from 2002-04. David has extensive experience in research management since he worked as a research manager and leading teams at Cambridge as well as at CSIRO in Australia. In mid-2012, David was appointed Director of the Australian Bioinformatics Network, an initiative of CSIRO, EMBL Australia and Bioplatforms Australia, which has helped strengthen the Australian bioinformatics community and been instrumental in the formation of ABACBS, the Australian Bioinformatics and Computational Biology Society.

Dr CX Chan

Dr Cheong Xin Chan, better known as CX, is a Senior Research Officer and a Great Barrier Reef Foundation Bioinformatics Fellow at the University of Queensland's Institute for Molecular Bioscience. CX's research interests include microbial genomics and evolution, scalable phylogenomics, and microbial ecology. CX has a BSc with First Class Honours in Industrial Biology from Universiti Teknologi Malaysia (2001), and an MPhil with distinction in Algal Biotechnology from the University of Malaya (2003). His earlier research work focused on functional genomics of tropical seaweeds. After obtaining his PhD in Genomics and Computational Biology from UQ in 2008, CX worked as postdoctoral associate with Professor Debashish Bhattacharya at Rutgers University (New Brunswick, New Jersey), where he worked primarily on algal genomics and evolution, and the origin of plastids. Currently CX is part of the Reef Future Genomics (ReFuGe) 2020 Consortium, working with Professor Mark Ragan at UQ in algal genome sequencing of the coral reef symbionts. While his research routinely involves de novo assembly and analysis of high-throughput sequencing data, CX also has a keen interest in developing highly scalable strategies in phylogenomics using advanced computational and database approaches. CX is an Associate Faculty Member of F1000 Prime (Genomics & Genetics, Bioinformatics), and currently a Teaching Fellow at UQ's School of Chemistry and Molecular Biosciences while supervising three PhD projects.

Dr Nic Waddell

Dr Nic Waddell is head of the Medical Genomics group and deputy coordinator of the Genetics and Computational Biology Department at QIMR Berghofer Medical Research Institute. She completed her PhD at the University of Leicester in 2003. She is an NHMRC Career Development Fellow, cancer researcher and bioinformatician who is an expert in the interpretation of multiple data types, including next generation sequence data. She is a member of the International Cancer Genome Consortium (ICGC) and the Australian Genomics Health Alliance (AGHA). Since 2010 she has received >$7 million as a Chief Investigator, and has published >50 papers including Nature (n=6), Nature Genetics (n=2) and AJGH (n=3). She leads the genomics of several cancer genome projects including mesothelioma and oesophageal cancer. Her research focuses on using next generation sequencing to address clinical challenges.

Dr Lynn Fink

Dr Lynn Fink leads a research group at The University of Queensland Diamantina Institute focused on using next-generation technologies to interrogate disease complexity and mechanisms controlling cancer progression in both blood cancers and solid tumours. Dr Lynn Fink's research interests are aimed at understanding intra-tumour heterogeneity and interactions with the microenvironment in multiple myeloma and how those factors contribute to pathogenesis and disease progression. She is also interested in using bioinformatic approaches to identify novel factors in epithelial plasticity and elucidate their role in cancer progression in solid tumours. Dr Lynn Fink obtained a Bachelor of Science degree from The University of Arizona with a research project on understanding melanomagenesis and the anti-melanoma activity of a potential chemotherapeutic peptide. Later on, Lynn obtained her Doctor of Philosophy degree from the University of California in the USA.

Dr Katherine Pillman

Dr Katherine Pillman is a Bioinformatics Research Fellow at the Centre for Cancer Biology at the University of South Australia. Katherine is a bioinformatician with a broad interest in many areas of gene regulation. She began her career as a wet-bench biologist, analysing gene expression regulatory networks in transgenic barley plants in Australia during her PhD, and then in potato plants in the USA during her first post-doc. Her experience with RNA-seq during this post-doc led her to change career direction to bioinformatics. In 2012, she returned to Australia to take up the role of lead bioinformatician with Prof Greg Goodall, working on gene regulatory mechanisms and networks in cancer. Their recent seminal paper identified the first protein known to control the formation of circular RNAs and characterising circular RNAs during the epithelial-to-mesenchymal transition. Her current work involves using a range of genomics next-generation sequencing data types to dissect gene regulation, including analysis of circular RNAs, alternative splicing, microRNA biology and targeting, epigenetic modifications, gene regulatory networks and expression.

AB^{3}ACBS CONFERENCE, DAY 1

Tuesday, 1 November

9.00am – 5.15pm, Level 5, P Block

Time	Activity
8:00-9:00am	AB3ACBS registration and poster hanging
9:00-10:30am	**SESSION 1:** Microbes
	• **Nouri Ben Zakour** Evolutionary epidemiology of successful bacterial pathogens • **Andrew Buultjens** Exploiting extremes of Legionella pneumophila genomic diversity for accurate source attribution • **Darryl Reeves** Probabilistic Inference and Shared Parameter Learning for Metagenomic Sequence Analysis • **Patrick Laffy** HoloVir: Taxonomic and functional analysis of viral metagenomic communities.
10:30-11:00am	Morning tea Level 5 Foyer
11:00-12:30pm	**SESSION 2:** Cancer
	• **Andreas Schreiber** Taking the confusion out of fusions: Structural mutation detection in cancer research and diagnostics • **Shila Ghazanfar** Exploring Pan-Cancer Network Relationships Between Somatic Changes and Expression Profiles with PACMEN • **Luis Lara-Gonzalez** Consequences of Drug Dose Modulation on Clonal Dynamics • **Rebecca C. Poulos** Functional mutations form at CTCF/cohesin binding sites in melanoma due to uneven nucleotide excision repair across the motif • **Christoffer Flensburg** Tracking clonal evolution in cancer from multiple samples
12:30-1:30pm	Lunch
1:30-3:00pm	**SESSION 3:** Statistics
	• **Matt Ritchie** Tools for comparing and combining RNA-seq results • **Paul Lin** CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-Seq data • **Belinda Phipson** Gene length bias in single cell RNA-seq data • **Joseph Cursons** Computational methods to examine micro-RNA targeting of interacting protein networks • **Göknur Giner** FRY: A Fast Approximation to ROAST Gene Set Test with Mean Aggregated Set Statistics
3:00-3:30pm	Afternoon tea
3:30-5:00pm	**SESSION 4:** Plants and Animals
	• **Kate Hertweck** Plant systematics to cancer biology: genome-wide patterns and organismal evolution • **Jimmy Breen** Whole methylome analysis of grapevine reveals tissue specific DNA methylation variation • **Andrew Lonsdale** Mighty Morphin FASTA Files • **Nadia Davidson** SuperTranscript: a compact reference for the transcriptome • **Stephen Fletcher** Small Complementary RnA Mapper (SCRAM): a tool for studying small interfering RNA biogenesis in plants
5:00pm	**Poster Session**
6:30pm	**Conference Dinner**

SPECIAL GUEST PRESENTATION

9:00am – 9:30pm, P514

EVOLUTIONARY EPIDEMIOLOGY OF SUCCESSFUL BACTERIAL PATHOGENS

Dr Nouri Ben Zakour

Principal research scientist
Westmead Institute for Medical Research

Abstract: The last 10 years have witnessed next-generation whole genome sequencing (WGS) drastically transform the fields of genomics, microbiology and epidemiology, among others. In this talk, I will explore how WGS provides invaluable insights into the pathogenesis, evolution and global dissemination of multi-drug resistant pathogens and helps us understand what it takes to be a successful bacterial pathogen. I will also touch on some of the new perspectives that this ongoing "genomics revolution" opens up for hospitals and clinical settings.

Biography: Dr Nouri Ben Zakour is a principal research scientist in microbial genomics at the Westmead Institute for Medical Research in Sydney. Prior to recently joining WIMR, she received her PhD in bioinformatics from the University of Rennes (France) and further specialised in pathogenomics at the University of Edinburgh and the University of Queensland. Her research currently focuses on the evolutionary epidemiology of established and emerging multi-drug resistant clinical and veterinary bacterial pathogens. In collaboration with global research, clinical and industrial leaders, she also aims at bringing microbial genomics closer with hospitals and veterinary settings by delivering applied solutions to outbreak tracking, antibiotic resistance surveillance and therapeutic targets identification.

SPECIAL GUEST PRESENTATION

11:00am – 11:30am, P514

TAKING THE CONFUSION OUT OF FUSIONS: STRUCTURAL MUTATION DETECTION IN CANCER RESEARCH AND DIAGNOSTICS

Dr Andreas Schrieber

Centre for Cancer Biology

Abstract: Next generation sequencing (NGS) has become a standard tool in cancer research. Used with DNA, NGS has enabled e.g. genome-wide identification of point mutations, structural rearrangements and transcription factor binding sites, while with RNA one can measure transcriptome-scale gene expression, splicing variation and discover deleterious fusion genes. From a bioinformatic perspective, mutation detection algorithms have matured enough to permit widespread uptake of NGS into diagnostic laboratories, including in our own institution. With RNASeq, however, uptake into a diagnostic setting is still stymied by changing wet-lab as well as bioinformatic techniques. In this talk, I will describe how developments in the wet-lab have caused us to reassess the bioinformatics of gene fusion detection and I will report on our efforts to improve the reliability of NGS gene fusion detection sufficiently to permit its uptake into diagnostics.

Biography: Andreas heads the bioinformatics group at the Centre for Cancer Biology's ACRF Cancer Genomics Facility. The group focuses on applied bioinformatics of high throughput experiments, ranging from analysis of transcriptomic, microarray or RNASeq data, gene regulation studies using ChIP and CLIPSeq, to the search for disease-associated point and structural mutations of the human genome. Embedded in the South Australian Department of Health, the facility and the bioinformatics group are closely involved with the implementation of NGS technologies into a diagnostic setting.

Andreas obtained a BSc and MSc from the University of Melbourne, followed by a PhD in theoretical nuclear/particle physics from the University of Adelaide in 1990. He completed postdocs in the Netherlands, Switzerland and Canada before returning to Australia on an ARC Research Fellowship. In 2002 he switched fields to Bioinformatics by joining the Australian Centre for Plant Functional Genomics at the Waite Campus of the University of Adelaide. He has been with the CCB since 2011.

SPECIAL GUEST PRESENTATION

1:30pm – 2:00pm, P514

TOOLS FOR COMPARING AND COMBINING RNA-SEQ RESULTS

Dr Matt Ritchie

Molecular Medicine

The Walter and Eliza Hall Institute of Medical Research

Abstract: In this talk I will discuss our recent efforts in RNA-mixology to design more realistic control experiments for benchmarking different RNA-seq analysis methods and protocols. I will also introduce new software for combining results from different gene set testing methods (EGSEA) and delivering results from RNA-seq analyses in a more interactive way (Glimma).

Biography: Dr Matt Ritchie is a statistical bioinformatician at the WEHI who develops analysis methods for RNA-seq data and other genome-wide approaches. His lab is currently exploring data from single cell technologies and long-read sequencing platforms to study transcription and methylation. He is a keen developer of open-source R/Bioconductor software for genomic analysis.

SPECIAL GUEST PRESENTATION

3.30pm – 4.00pm, P514

PLANT SYSTEMATICS TO CANCER BIOLOGY: GENOME-WIDE PATTERNS AND ORGANISMAL EVOLUTION

Associate Professor Kate Herweck
University of Texas

Abstract: Obtaining data is no longer the limiting factor in genomic research. While such data offer incredible promise, we are faced with increasing demands to store, manage, and extract meaning from available genomes. Reproducible science skills allow researchers to meet these needs as well as associate genome-wide data with organismal information, like life history traits and phenotypes. Bioinformatics, therefore, offers unprecedented opportunities to reveal patterns in both evolutionary history and future trajectories of organisms. Three seemingly disparate but complementary examples will highlight the union of genome and organismal evolution: molecular systematics in monocotyledonous plants, transposable element proliferation in experimentally evolved Drosophila, and somatic mutation in cancer biology. Projects integrating reproducible science with evolutionary biology provide invaluable opportunities for student training, preparing the next generation of scientists to synthesize patterns across genomes.

Biography: Kate L. Hertweck is an Assistant Professor in the Department of Biology at the University of Texas at Tyler specializing in genomics, bioinformatics, and evolutionary biology. She received her B.S. in Biology from Western Kentucky University and Ph.D. in Biology from University of Missouri, where she studied systematics and evolution in plants. She traded in her lab coat and field boots for full-time computational work by joining the National Evolutionary Synthesis Center (NESCent, at Duke University) as a postdoctoral fellow, where she had the opportunity to refine her skills in data analysis and computational biology. Her current research spans the full breadth of genomic analysis, from populations through deep evolutionary time and humans to non-model systems.

SESSION 1: BRIEF PRESENTATIONS

3 x 15 minute talks (12 mins + 3 mins for questions)

Paper	Presenting author
Exploiting extremes of Legionella pneumophila genomic diversity for accurate source attribution	*Andrew Buultjens* Department of Microbiology and Immunology at the Peter Doherty Institute for Infection and Immunity, The University of Melbourne, Victoria, Australia
Probabilistic Inference and Shared Parameter Learning for Metagenomic Sequence Analysis	*Darryl Reeves* Department of Physiology and Biophysics, Weill Cornell Medical College, New York, NY, USA
HoloVir: Taxonomic and functional analysis of viral metagenomic communities.	*Patrick Laffy* Australian Institute of Marine Science, Townsville, Queensland, Australia

ORAL PRESENTATION ABSTRACTS

Exploiting extremes of Legionella pneumophila genomic diversity for accurate source attribution

Andrew H. Buultjens (a), (b), Kyra Y. L. Chua (c), Sarah L. Baines (a), Jason Kwong (a), (b), (c), Wei Gao (b), Mark B. Schultz (a), (c), Zoe Cutcher (d), (e), Stuart Adcock (d), Susan Ballard (a), Takehiro Tomita (a), Nela Subasinghe (a), Glen Carter (b), (c), Sacha J. Pidot (b), (c), Lucinda Franklin (d), Torsten Seemann (c), (f), Anders Gonçalves Da Silva (a), (c), Benjamin P. Howden (a), (b), (c), Timothy P. Stinear (b), (c)

(a) Department of Microbiology and Immunology at the Peter Doherty Institute for Infection and Immunity, The University of Melbourne, Victoria, Australia
(b) Doherty Applied Microbial Genomics, The Peter Doherty Institute for Infection and Immunity, Victoria, Australia
(c) Microbiological Diagnostic Unit Public Health Laboratory at the Peter Doherty Institute for Infection and Immunity, The University of Melbourne, Victoria, Australia.
(d) Health Protection Branch, Department of Health and Human Services, Victoria, Australia
(e) National Centre for Epidemiology and Population Health, Australian National University, Canberra, Australia
(f) Victorian Life Sciences Computational Initiative, The University of Melbourne, Victoria, Australia

Public health agencies are increasingly using genomics in Legionnaires' disease investigations, relying on similarities in the core genomes of the causative bacteria (Legionella pneumophila) to identify environmental contamination sources. Here, we show that the assumptions underlying these studies are flawed. We propose instead a statistical learning approach for source attribution that also considers accessory genome variation. We compared genomes of 234 L. pneumophila isolates obtained from patients and cooling towers in Melbourne, Australia between 1994 and 2014. This collection spanned 29 infection clusters, including one of the largest reported Legionnaires' disease outbreaks, involving 125 cases at an aquarium. There was one dominant genotype that exhibited startlingly low core genome variation. The median pairwise nucleotide difference for the 180 genomes obtained across Melbourne over 21 years was only 5 single nucleotide polymorphisms (SNPs) (IQR 3-7). In addition to high sequence conservation, we also uncovered within-outbreak isolate diversity. By assessing only cooling tower isolates and including all genomic variation (SNPs and accessory genome), we built a multivariate model to find cooling tower-specific genomic signatures. We then used this model to accurately predict the origin of clinical isolates. A sister model built with SNPs from the recently advocated cgMLST approach applied to this same population was poorly predictive. These data show that health agencies will require a deep understanding of local L. pneumophila population structure or risk source misattribution. The restricted genetic diversity seen here also suggests environmental reservoirs of quiescent bacteria sporadically seeding warm water sources, causing human cases of Legionellosis.

Probabilistic Inference and Shared Parameter Learning for Metagenomic Sequence Analysis

Darryl Reeves (a, b, c) and Christopher E. Mason (a,b,c,d)

(a) Department of Physiology and Biophysics, Weill Cornell Medical College, New York, NY, USA
(b) The HRH Prince Alwaleed Bin Talal Bin Abdulaziz Alsaud Institute for Computational Biomedicine, New York, NY, USA
(c) Tri-Institutional Training Program in Computational Biology and Medicine, New York, NY, USA
(d) The Feil Family Brain and Mind Research Institute (BMRI), New York, NY, USA

Methods utilized to understand the composition of organisms in metagenomic sequencing data have applications in a number of different areas relevant to human health and disease. The Human Microbiome Project (HMP) introduced a baseline understanding of the underlying composition of organisms that make up the microbial environment of different regions of the human body in healthy individuals [1]. This ongoing research is relevant in a number of disease-related research efforts in which the human microbiome is implicated in the development or protection from disease [2, 3]. In addition, environmental metagenomic sequencing efforts have been undertaken to understand how external microbial environments can affect human health [4]. Some of these microbial communities are simply interesting in their own right, as they can help to further our understanding of the ability for organisms to adapt to extreme environments [5]. One of the challenges with metagenomic sequencing data is identifying and estimating the abundance of organisms in a sequencing sample given short read sequences and an incomplete set of reference genomes.

Our method, provides a novel approach to this problem in 3 ways: 1) By leveraging a probabilistic inference engine, sequencing data is analysed as a whole in addition to the read-by-read analysis approach used in many previous methods. This approach helps to avoid misclassifications that are made when a read matches a reference sequence closely but has low probability of being a match in the context of the entire sequencing sample. 2) By utilizing the KMerge [6] k-mer hashing framework, the computational memory footprint of a reference sequence can be reasonably bounded regardless of the genome length. This allows metagenomic sequencing identification and abundance estimation to be performed with a larger set of reference genomes across kingdoms. 3) While originally designed for high-throughput shotgun sequencing data, the method is designed in such a way that data from 16S rRNA studies can also be analysed by simply changing the set of reference sequences used in the analysis. Such a flexible approach allows for diverse datasets from the same environment to be analysed using a unified approach.

[1] Huttenhower et al. (2012). Nature, 486(7402), 207–214.

[2] Proal et al. (2013). Current Opinion in Rheumatology, 25(2), 234–240.

[3] Parekh et al. (2015). Clin Trans Gastroenterol, 6, e91.

[4] Casas et al. (2016).Current Environmental Health Reports, 3(3), 238–249.

[5] Bragina et al. (2014). Molecular Ecology, 23(18), 4498–4510.

[6] Reeves et al. (2016). KMerge: A computational frame for genome comparisons using hashed k-mer frequencies. Manuscript in preparation.

HoloVir: Taxonomic and functional analysis of viral metagenomic communities.

Patrick W Laffy(a), Elisha M Wood-Charlson(b), Dmitrij Turaev(c), Karen Weynberg(a), Emmanuelle Botte(a), Madeleine van Oppen(a,d), Nicole S Webster(a), Thomas Rattei(c)

(a) Australian Institute of Marine Science, Townsville, Queensland, Australia.
(b) Center for Microbial Oceanography: Research and Education (C-MORE), University of Hawai'i at Mānoa, Honolulu, HI, USA.
(c) Department of Microbiology and Ecosystem Science, Division of Computational Systems Biology, University of Vienna, Vienna, Austria.
(d) School of Biosciences, University of Melbourne, Parkville, Australia

Whilst abundant bioinformatics resources have been developed to analyse the taxonomic and functional composition of microbial metagenomic data, their applicability to viral metagenomics is limited. HoloVir is a computational workflow designed to process large viral metagenomics datasets, enabling thorough taxonomic and functional characterisation of viral communities. HoloVir performs taxonomic assignment via pairwise comparsions to the Viral Refseq genomic resource, and incorporates marker analysis to identify potential cellular contaminants and estimate viral taxonomic composition. HoloVir incorporates single read analysis, combined metagenome assembly and gene prediction, and has been validated using simulated viral community datasets. Broad functional classification of predicted genes is performed by assigning COG microbial functional categories using EggNOG, and higher resolution functional analysis is performed using SwissProt keyword assignment. The workflow has been shown to be flexible enough to accommodate taxonomically diverse hosts, yet specific enough to identify differences within the associated viral assemblages. HoloVir has been successfully applied to investigate viral communities within several key holobionts from the Great Barrier Reef and has facilitated the comparison of viral communities across species, health states and life history stages. A recent addition of GRASPx analysis of viral orthologous groups has facilitated the analysis of viral strain diversity in metagenomes, uncovering the importance of ssDNA virus strain diversity in coral and sponge virome composition. This advance in microdiversity identification is essential for understanding host-virus interactions, and will be incorporated into a future release of HoloVir. The HoloVir workflow is available to download via GitHub (https://github.com/plaffy/HoloVir).

SESSION 2: BRIEF PRESENTATIONS

4 x 15 minute talks (12 mins + 3 mins for questions)

Paper	Presenting author
Exploring Pan-Cancer Network Relationships Between Somatic Changes and Expression Profiles with PACMEN	*Shila Ghazanfar* The University of Sydney
Consequences of Drug Dose Modulation on Clonal Dynamics	*Luis Lara-Gonzalez* Bioinformatics & Cancer Genomics Laboratory, Peter MacCallum Cancer Centre, Melbourne
Functional mutations form at CTCF/cohesin binding sites in melanoma due to uneven nucleotide excision repair across the motif	*Rebecca C. Poulos* Prince of Wales Clinical School and Lowy Cancer Research Centre, Sydney, Australia
Tracking clonal evolution in cancer from multiple samples	*Christoffer Flensburg* WEHI, Melbourne, Australia

ORAL PRESENTATION ABSTRACTS

Exploring Pan-Cancer Network Relationships Between Somatic Changes and Expression Profiles with PACMEN

Shila Ghazanfar (a,b), Jean Yee Hwa Yang (a)

(a) The University of Sydney; (b) CSIRO

The Cancer Genome Atlas is a rich source of information enabling study across and between cancers. Recently, network approaches have been applied to such data to uncover complex interrelationships between somatic changes and expression profiles, but lack direct testing. In this pan-cancer study we co-analysed somatic changes (mutation and copy number alterations) and expression (gene or protein) to identify networks of interest, shedding light on these relationships in a direct manner.

Using somatic changes and gene expression information across each of 19 cancers, we identified mutation-expression networks and enabled interrogation through an online interactive R Shiny application, PAn Cancer Mutation Expression Networks (PACMEN). Analysis involved directly testing for differences in expression abundance via somatic changes. PACMEN allows data and parameter choice, showcases analyses performed using curated and estimated networks, and provides network description and visualisation.

We found networks identified were significantly enriched for known cancer-related genes such as melanoma ($P<0.01$ Network of Cancer Genes 4.0). Notably, comparison between cancers showed a greater overlap of nodes for cancers with higher overall mutation load (melanoma, lung), compared to those with a lower overall mutation load (glioblastoma, leukemia).

We propose and implement a framework for exploring network information through coanalysis of somatic changes and gene/protein expression profiles. Our pan-cancer approach suggests that while mutations are frequently common among cancer types, the impact they have on surrounding networks via expression changes varies, which may explain differences in efficacy of therapies among different diseases. In contrast, for some cancers mutation-associated network behaviour appears similar, suggesting a framework for uncovering cancers where similar therapeutic strategies may be applicable. PACMEN is available at <shiny.maths.usyd.edu.au/PACMEN>.

Consequences of Drug Dose Modulation on Clonal Dynamics

Luis Lara-Gonzalez(a), David Goode (a,b), Sherene Loi (b,c), Davide Ferrari (d), Anthony Papenfuss (a,e)

(a) Bioinformatics & Cancer Genomics Laboratory, Peter MacCallum Cancer Centre, Melbourne, Victoria 3000, Australia
(b) Sir Peter MacCallum Department of Oncology, The University of Melbourne, Victoria 3010, Australia
(c) Translational Breast Cancer Laboratory, Peter MacCallum Cancer Centre
(d) Mathematical and Computational Biology, Statistics, The University of Melbourne, Victoria 3010, Australia
(e) Computational Biology, Bioinformatics Division, Walter and Eliza Hall, Victoria 3052, Australia

By reconstructing tumour evolution, computational modelling is making significant progress toward identifying drug resistance origins and optimal drug timing and order. However, data-driven assessment is lacking. We are creating an onco-bioinformatic tool to study the dynamics of tumour growth, treatment, and resistance to improve our understanding of cancer and the design of better treatments for the disease.

We modeled tumour evolution as an agent-based discrete time branching process that tracks the expansion of diverse clonal lineages as they acquire driver and passenger mutations that alter their proliferation and mutation rates. Clonal proliferation is subject to a spatio-temporal size-dependent penalty to provide characteristic tumour growth patterns. Once the tumour attains a diagnosable size (1 to 4 billion cells), a mitotic phase-specific perturbation is introduced to model anticancer agents. This environmental disruption impacts clonal dynamics, and we observed diverse heterogeneity, genomic instability, and resistance evolutionary paths. Our tool recovers various tumour development rates seen in the clinic, in which genomic instability promotes clonal diversification, leading to a state of invasiveness and prevailing (cross)-resistance.

Our tool predicts that the last couple of years prior to diagnosis are essential in the pathogenesis of the tumour, which requires 2 - 6 driver mutations to bypass the effects of anticancer agents. Our simulated clinical trials comparing cytotoxic and targeted drug combinations show that moderate-dose schemes lead to prolonged survival rates, even in the presence of pre-existing drug resistant clones. Therefore, treatments maintaining clonal proportions should be considered as an alternative way for tumour growth control.

Functional mutations form at CTCF/cohesin binding sites in melanoma due to uneven nucleotide excision repair across the motif

Rebecca C. Poulos (a), Julie A. I. Thoms (a), Yi Fang Guan (a), Ashwin Unnikrishnan (a), John E. Pimanda (a,b) and Jason W. H. Wong (a)

(a) Prince of Wales Clinical School and Lowy Cancer Research Centre, UNSW Australia, Sydney NSW 2052, Australia
(b) Department of Haematology, Prince of Wales Hospital, Sydney NSW 2052, Australia

CCCTC-binding factor (CTCF) is a crucial protein involved in maintaining the three-dimensional organisation of the genome and defining regions of gene expression. CTCF binding sites are frequently mutated in cancer, but how these mutations accumulate and whether they broadly perturb CTCF binding is not well understood. In this study, we analysed the genomes of 52 skin cancer samples and found skin cancers to exhibit a unique and asymmetric mutation pattern within CTCF motifs. To understand the mechanisms underlying mutation formation, we examined datasets of nucleotide excision repair (NER), CTCF binding and replication timing, showing the mutation pattern to be attributable to ultraviolet irradiation and differential NER across individual nucleotides within CTCF motifs. We additionally demonstrated that CTCF binding site mutations form independent of replication timing, with mutations enriched at sites of CTCF/cohesin complex binding and suggesting a novel role for cohesin in stabilizing CTCF-DNA binding and impairing NER in a multifactorial manner. To demonstrate that CTCF binding site mutations are functional in melanoma, we performed CTCF ChIP-seq in a melanoma cell-line, reporting allele-specific reduction of CTCF binding to mutant alleles. Investigating whether selection underlies CTCF motif mutation accumulation, we analysed topologically-associated domains in skin cells and identified mutated CTCF anchors which contain differentially-expressed cancer-associated genes. One gene identified was the tumour suppressor APC, a key component of the Wnt signaling transduction pathway and a gene previously implicated in melanoma development. Through bootstrapping analyses and mutation simulation however, we found that genome-wide, CTCF motif mutations in melanoma are generally under neutral selection. Regardless, the frequency and potential functional impact of such mutations highlights the need to consider their impact on cellular phenotype in personalized genomes.

Tracking clonal evolution in cancer from multiple samples

Christoffer Flensburg, Ian Majewski

WEHI, Melbourne, Australia

Cancer is constantly evolving. To understand the disease, we need to monitor how it changes, and the advent of genome wide DNA and RNA sequencing provides a powerful way of doing this. Our lab has developed methods to take full advantage of sequencing data from multiple cancer samples from a single individual. This allows us to track the clonal evolution of a cancer: pinpointing molecular changes in cancer cells that resist therapy, spread around the body or transform into more aggressive diseases. I will illustrate some of these methods through a range of examples.

We identify clonal populations based on somatic mutations, both single nucleotide variants and copy number alterations. In some individuals we see genes or pathways being repeatedly mutated in distinct cell population, which can identify driver mutations from small cohorts, or even from a single individual. We see the same copy number alteration present in multiple samples affect different alleles. This shows that the copy number alterations in the two samples are different events, which again points towards a driver mutation. Clonal consistency constraints allow us to filter noise and inform the choice of phylogenetic tree of the cancer. We combine captured DNA and RNA from the same sample to identify cis acting regulatory mutations that could not have been found from the DNA or RNA alone.

In combination, these methods enable significant findings from multiple cancer samples that would not be possible from a naïve cancer-normal analysis.

SESSION 3: BRIEF PRESENTATIONS

4 x 15 minute talks (12 mins + 3 mins for questions)

Paper	Presenting author
CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-Seq data	*Paul Lin* Victor Chang Cardiac Research Institute
Gene length bias in single cell RNA-seq data	*Belinda Phipson* Murdoch Childrens Research Institute
Computational methods to examine micro-RNA targeting of interacting protein networks	*Joseph Cursons* Walter and Eliza Hall Institute of Medical Research
FRY: A Fast Approximation to ROAST Gene Set Test with Mean Aggregated Set Statistics	*Göknur Giner* Walter and Eliza Hall Institute of Medical Research

ORAL PRESENTATION ABSTRACTS

CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-Seq data

Paul Lin

Victor Chang Cardiac Research Institute

Most existing dimensionality reduction and clustering packages for single-cell RNA-Seq (scRNA-Seq) data deal with dropouts by heavy modelling and computational machinery. Here we introduce CIDR (Clustering through Imputation and Dimensionality Reduction), an ultrafast algorithm which uses a novel yet very simple 'implicit imputation' approach to alleviate the impact of dropouts in scRNA-Seq data in a principled manner. Using a range of simulated and real data, we have shown that CIDR outperforms the state-of-the-art methods, namely t-SNE, ZIFA and RaceID, by at least 50% in terms of clustering accuracy, and typically completes within seconds for processing a dataset of hundreds of cells. CIDR can be downloaded at https://github.org/VCCRI/CIDR.

Gene length bias in single cell RNA-seq data

Belinda Phipson (a), Luke Zappia (a,b) and Alicia Oshlack (a,b)

(a) Murdoch Childrens Research Institute
(b) School of Biosciences, University of Melbourne

Single cell RNA sequencing (scRNA-seq) has rapidly gained popularity for profiling transcriptomes of hundreds to thousands of single cells. This new technology has led to the discovery of novel cell types and revealed insights into the development of complex tissues. However, many technical challenges need to be overcome during data generation. Due to minute amounts of starting material, samples undergo extensive amplification, which increases technical variability. A solution for mitigating amplification biases is to include Unique Molecular Identifiers (UMIs), which tag individual molecules. Transcript abundances are then estimated from the number of unique UMIs aligning to a specific gene and PCR duplicates resulting in copies of the UMI are not included in expression estimates.

Gene length bias is well understood in bulk RNA-seq data. When cDNAs are fragmented long genes result in more fragments for the same number of transcripts, resulting in higher counts and more power to detect differential expression. As a result gene set testing is biased towards categories containing longer genes. Here we investigate the effect of gene length bias in scRNA-seq across a variety of technical protocols. As hypothesised, we find that scRNA-seq datasets that have been sequenced using a full-length transcript protocol exhibit gene length bias such that shorter genes tend to have lower counts and a higher rate of dropout. In contrast, protocols that include UMIs do not exhibit gene length bias. In addition, UMI protocols reveal that shorter genes are as highly expressed as longer genes, and dropout is mostly uniform across genes of varying length. This result influences down-stream analysis such as gene set testing which will vary based on the protocol used. In particular, care should be taken when gene abundances are estimated in UMI data, as naively calculating RPKMs by dividing by gene length will artificially inflate the expression of shorter genes relative to longer genes.

Computational methods to examine micro-RNA targeting of interacting protein networks

Joseph Cursons and Melissa J Davis

Division of Bioinformatics, Walter and Eliza Hall Institute of Medical Research

Micro-RNAs (miRs) are small regulatory transcripts, approximately 22 nucleotides in length, which exert post transcriptional control over other RNA transcripts. Targeting of miRs is mediated through Watson-Crick sequence complement, generally over 6-8 aligned nucleotides within the 3’ UTR of target mRNAs.

Micro-RNAs have been shown to bind and regulate large numbers of targets, both experimentally, and through computational approaches that examine sequence overlap. Given this result, miRs are often thought to be non-specific. Using network-based computational methods which examine micro-RNA targeting in the context of protein-protein interactions, however, it can be shown that many micro-RNAs which are co-regulated across disease states target mRNA transcripts that encode proteins with functional interactions. Thus, although miRs are promiscuous in their binding, they are specific in their targeting of defined molecular regulatory programs.

I will describe the computational methods used for this work, including data sources, and the miR-target prediction databases DIANA-microT and TargetScan. The python library networkx will be discussed as an approach to work with network structures, and I will show how it can be coupled with matplotlib to produce annotated graphical network structures for aiding interpretation. Finally, I will introduce permutation testing as a method to examine the significance of observed network structures/features in a rigorous, quantitative manner.

FRY: A Fast Approximation to ROAST Gene Set Test with Mean Aggregated Set Statistics

Göknur Giner (a,b) and Gordon K. Smyth (a,c)

(a) Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville VIC 3052, Australia
(b) Department of Medical Biology, The University of Melbourne, Parkville VIC 3010, Australia
(c) Department of Mathematics and Statistics, The University of Melbourne, Parkville VIC 3052, Australia

Gene set tests are often used in differential expression analyses to explore the behaviour of a group of related genes. This is useful for identifying large-scale co-regulation of genes belonging to the same biological process or molecular pathway.

One of the most flexible and powerful gene set tests is the ROAST method in the limma R package. ROAST uses residual space rotation as a sort of continuous version of sample permutation. Like permutation tests, it protects against false positives caused by correlations between genes in the set. Unlike permutation tests, it can be used with complex experimental design and with small numbers of replicates. It is the only gene set test method that is able to analyse complex “gene expression signatures” that incorporate information about both up and down regulated genes simultaneously.

ROAST works well for individual expression signatures but has limitations when applied to large collections of gene sets, such as the Broad Institute’s Molecular Signature Database with over 8000 gene sets. In particular, the p-value resolution is limited by the number of rotations that are done for each set. This makes it impossible to obtain very small p-values and hence to distinguish the top-ranking pathways from a large collection. As with permutation tests, the p-values for each set may vary from run to run.

This talk presents Fry, a very fast approximation to the complete ROAST method. Fry approximates the limiting p-value that would be obtained by performing a very large number of rotations with ROAST. Fry preserves most of the advantages of ROAST but also provides exact high-resolution p-values very quickly. In particular, it can distinguish the most significant sets in large collections and to yield statistically significant results after adjustment for multiple testing. This makes it an ideal tool for large-scale pathway analysis.

Another important consideration in gene set tests is the possible unbiased or incorrect estimation of P-values due to the correlation among genes in the same set or dependence structure between different sets.

Children's Medical Research Institute Jeans for Genes
The Westmead Institute for Medical Research
The Sydney children's Hospitals Network
Westmead Hospital
kids research institute the children's hospital at Westmead
NSW Government Health Western Sydney Local Health District
The University of Sydney

SESSION 4: BRIEF PRESENTATIONS

4 x 15 min talks (12 mins + 3 mins for questions)
3:30pm – 5:00pm, P514

Paper	Presenting author
Whole methylome analysis of grapevine reveals tissue specific DNA methylation variation	*Jimmy Breen* Robinson Research Institute, University of Adelaide
Mighty Morphin FASTA Files	*Andrew Lonsdale* ARC Centre of Excellence in Plant Cell Walls, University of Melbourne
SuperTranscript: a compact reference for the transcriptome	*Nadia Davidson* Murdoch Childrens Research Institute
Small Complementary RnA Mapper (SCRAM): a tool for studying small interfering RNA biogenesis in plants	*Stephen Fletcher* School of Chemistry and Molecular Biosciences, The University of Queensland

Whole methylome analysis of grapevine (Vitis vinifera) reveals tissue specific DNA methylation variation

Breen J (a,b), Rodriguez Lopez CM (b), De Bei R (b), David R (c), Searle I (d), Collins C (c)

(a) Robinson Research Institute, University of Adelaide, Adelaide SA
(b) Environmental Epigenetics and Genetics Group, School of Agriculture, Food and Wine, University of Adelaide, Adelaide SA
(c) School of Agriculture, Food and Wine, University of Adelaide, Adelaide SA
(d) School of Biological Sciences, University of Adelaide, Adelaide SA

Wine made from grapevine (Vitis vinifera) vineyards is an important agricultural export for the national economy, and a crucial industry for South Australia. The quality of wine is known to be extensively influenced by numerous environmental factors such as climate, altitude and soil, a French concept known as “terroir” (meaning earth and land). Along with the suspected influence of the environment on the quality of wine, grapevine is a clonally propagated crop, which limits the possible genetic influence on wine quality, reinforcing the supposed role of epigenetics (functionally relevant changes to the genome that do not involve a change in the nucleotide sequence). In this study, we investigate the role of DNA methylation on differentiation of three grapevine organs (berry skin, seed and leaf), across two berry development stages key to wine quality (Veraison and Harvest), at one individual site by producing the first single base 5’ methyl-cytosine methylomes for Vitis vinifera. After confirming the lack of genetic variation between sampled individuals using 5.7 million SNPs from whole genome resequencing data, we identify tissue and development specific DNA methylation. Our results are consistent with those reported for other angiosperm methylomes (such as Brassicaceae species etc), especially when comparing average CpG (64%), CHG (45%) and CHH (8%) methylation. In a genomic context, no significant changes on DNA methylation density were at the transcription start sites of genes, however significant variation was found at the start sites of transposable elements, especially in seed tissues. We suggest that changes in methylation patterns seen in seeds to be reflective of the mixture of early developmental tissue compared to tissues that contain cell types from later developmental stages (such as leaf and berry skins), reflecting age-related accumulation of DNA methylation. In all, this data is an important profile of DNA methylation in grapevine, as well as an important tool in studying the epigenetic effects of age in clonally propagated fruit crops.

Mighty Morphin FASTA Files

Andrew Lonsdale

ARC Centre of Excellence in Plant Cell Walls, School of BioSciences, The University of Melbourne, Parkville, VIC, Australia,

How reliable is bioinformatics software? A challenge in determining the reliability and accuracy of many tools is that in the search for new knowledge, a 'gold standard' to test against does often not exist. Metamorphic testing can be used to overcome this issue by comparing the output of bioinformatics programs when input data is deliberately altered, as shown in recent work by Giannoulatou et al. [1]. Approaches inspired by metamorphic testing can reveal errors and biases by discovering unexpected changes in program output [2].

To encourage the use of metamorphic testing in bioinformatics, Mighty Morphin FASTA Files (MMFF) is an open source Python script and an accompanying web interface for generating test data in the common FASTA sequence file format. It is intended to complement other forms of testing on programs that use the FASTA file format. Various metamorphic relationships (MRs) can be applied to a seed of test data to produce additional FASTA files for each MR. Changes to the file format, sequence header, sequence content (nucleotides or amino acids), and sequence order can each be applied, and the results compared to unmodified data.

MMFF can also produce other classes of test data such as boundary and coverage testing. It is a simple framework for quickly generating test data that can then be run through existing programs and pipelines or during the development of new programs. Mighty Morphin FASTA Files (MMFF) can be used to improve and validate bioinformatics software. Silly name, serious purpose.

[1] Eleni Giannoulatou, Shin-Ho Park, David T Humphreys and Joshua WK Ho (2014). "Verification and validation of bioinformatics software without a gold standard: a case study of BWA and Bowtie". DOI: 10.1186/1471-2105-15-S16-S15

[2] Lonsdale, Andrew, Melissa J. Davis, Monika S. Doblin, and Antony Bacic (2016). "Better Than Nothing? Limitations of the Prediction Tool SecretomeP in the Search for Leaderless Secretory Proteins (LSPs) in Plants. DOI: 10.3389/fpls.2016.01451.

SuperTranscript: a compact reference for the transcriptome

Nadia M Davidson (a), Anthony DK Hawkins (a) and Alicia Oshlack (a,b)

(a) Murdoch Childrens Research Institute
(b) School of BioSciences, University of Melbourne

Profiling the transcriptome through RNA-Sequencing (RNA-Seq) is now common place, however visualising and analysing its complexity remains challenging. In particular, for species where no reference is available, many analysis approaches are impossible. A standard RNA-seq analysis for these non-model organisms begins with the de novo assembly of the reads into transcripts. However the assembled transcriptome necessarily contains repeated sequence because different isoforms of a gene share exons. This repeated sequence limits the use of many of the visualisation and analysis approaches which were originally developed for species with a reference genome.

In order to resolve these limitation we developed the concept of a superTranscript, a single sequence representation for each gene, consisting of all the unique exonic sequence, in transcriptional order. We have developed software, Lace, to construct superTranscripts from any set of transcripts. We show how using superTranscripts as a reference in non-model organisms allows for the first time, visualisation of read coverage across a gene, detection of differential isoform usage and variant calling. In addition we demonstrate the benefit of using SuperTranscripts in the RNA-Seq analyses of well annotated organisms. For example, a reference and de novo assembled transcriptome can be combined into a compact superTranscriptome, allowing easy identification of novel transcribed sequence. As an example we created superTranscripts from a chicken RNA-seq data set and discovered conserved coding sequence in over 1000 genes that was missed in the reference genome.

Small Complementary RnA Mapper (SCRAM): a tool for studying small interfering RNA biogenesis in plants

Fletcher SJ (a,b), Mitter N (b) and Carroll BJ (a)

(a) School of Chemistry and Molecular Biosciences, The University of Queensland, Brisbane, QLD, 4072, Australia
(b) Queensland Alliance for Agriculture and Food Innovation, The University of Queensland, Brisbane, QLD, 4072, Australia

In plants, small RNAs play key roles in the regulation of gene expression, defence against pathogens, and maintenance of genome stability via repression of transposable elements. Small RNAs function at the transcriptional level by directing genomic regions for chromatin modification, and at the post-transcriptional level by targeting transcripts for degradation.

To characterise the function of specific categories of small RNAs, as well as components in their biogenesis pathways, we have developed the Small Complementary RnA Mapper (SCRAM - http://carroll-lab.github.io/scram/), a fast, simple, lightweight aligner that generates publication-quality images from next-generation sequence data. The primary focus of this package is the alignment of small interfering RNAs (siRNAs), a class of small RNAs generated from long double-stranded RNA (dsRNA) templates, often viral replicative intermediates. As these siRNAs are derived from and are complementary to their target sequence, only the simplest of alignment algorithms is required. Accordingly, SCRAM rapidly aligns collapsed reads of a discrete length with a single command, and generates two varieties of plots: smoothed alignment profiles for a single reference sequence, and scatter plots of alignment abundance to multiple reference sequences, such as a cDNA or transposable element datasets, for the comparison of two experimental conditions. Both plot styles allow for unambiguous visual interpretation of small RNA alignments, making the output ideal for presentations and journal publications.

To date, SCRAM has been used in a number of studies, ranging from investigating the profound changes in siRNA biogenesis in peanut in response to Tomato spotted wilt virus infection, to elucidating the crucial role of DCL2, a Dicer-like enzyme that cleaves dsRNA into 22 nt siRNAs, in systemic post-transcriptional silencing in plants. The use of SCRAM as an important component in these studies will be presented.

AB^{3}ACBS FAST FORWARD PRESENTATIONS

Nov 1st, 5pm – The Cube

	Presentation
1	The Network Basis of Negative Genetic Interactions in Saccharomyces cerevisiae **Ignatius Pang,** *University of New South Wales*
2	Quantification of paediatric burn blister fluid proteome using SWATH MS to assist better clinical diagnosis **Tuo Zang,** *Queensland University of Technology*
3	A transcriptional signature for TGFβ-induced epithelial-mesenchymal transition in cancer **Momeneh Foroutan**, *The University of Melbourne*
4	Genetically Perturbed Pathways and Core Driver Genes in Late-Onset Alzheimer's Disease **Song Gao** South Australian Health & Medical Research Institute, Adelaide
5	Culture independent genome sequencing provides new insight into the microbiome associated with different types of diabetic foot ulcers (DFUs) **Sumeet Sandhu**, *Queensland University of Technology*
6	Outbreak of carbapenem-resistant Acinetobacter baumanii (CRAB) in a Brisbane Intensive Care Unit **Leah Roberts,** *The University of Queensland*
7	Effect of Serum Concentration on the Proteome of Rat Bone Marrow-Derived Mesenchymal Stem Cells **Morgan Carlton**, *Queensland University of Technology*
8	Comprehensive evaluation of the molecular and cellular activity of therapeutic small molecule BET inhibitors **Enid Lam**, *University of Melbourne*
9	Biomarker selection incorporating data independent acquisition mass spectrometry for the prediction of response to treatment in a non-healing wound cohort **Daniel Broszczak**, *Queensland University of Technology*
10	InsituNet, a Cytoscape app for network visualisation of in situ sequencing data **John Salamon**, *South Australian Health and Medical Research Institute, Adelaide*

Wednesday 2 November

Time	Activity
9:00-10:30am	**SESSION 5:** Evolution
	• **Simon Ho** Phylogenomic analysis and molecular evolutionary clocks • **Åsa Pérez-Bercoff** Investigating the evolution of new biochemical pathways in baker's yeast Saccharomyces cerevisiae • **Miles Benton** Identification of allelic-specific methylation profiles across generations in the Norfolk Island genetic isolate • **Kevin Murray** Estimating genetic similarity with the k-mer Weighted Inner Product (kWIP) • **Martin Smith** De novo characterisation of RNA structure motifs from ENCODE RIPseq data
10:30-11:00am	Morning tea
11:00-12:30pm	**SESSION 6:** ‘Omics
	• **Ute Roessner** Metabolomics - an important piece in the 'omics puzzle • **Pip Griffin** EMBL Australia Bioinformatics Resource (EMBL-ABR) • **Jovana Maksimovic** Through the looking glass: using Monocle to visualise methylation array data • **Ruth Fuhrman-Luck** Big data from a little proteolysis: Combining multiple omics platforms to identify novel functions of a protease associated with prostate cancer • **Lawrence Buckingham** Geometric map-reduce algorithms for alignment-free sequence comparison
12:30-1:00pm	Lunch
1:00-2:00pm	ABACBS Annual General Meeting
2:00-3:40pm	**SESSION 7:** Disease
	• COMBINE Best Student Talk • **Daniel Cameron** GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly • **Harriet Dashnow** Detecting pathogenic STR expansions in next-gen sequencing data • **Rod Lea** BootNet: a bootstrapping application for GLMnet modelling to identify robust classifiers in genomics data • **Alexis Lucattini** Assessing the Practicality of Oxford Nanopore Sequencing in Clinical Diagnositcs
3:40-5:00pm	**SESSION 8**: Systems
	• **Anna Trigos** How do vulnerabilities left during the evolution of cellular networks set the stage for cancer? • **Stuart Archer** TCP-seq, a novel technique for investigating mechanisms and regulation of eukaryotic translation initiation • **Teresa Attwood** The Road to Utopia: challenges in linking literature & research data
5:00pm	**Close**

SPECIAL GUEST PRESENTATION

9:00am – 9:30am, P514

Phylogenomic analysis and molecular clocks

Professor Simon Ho

Professor of Molecular Evolution
University of Sydney

Abstract: Evolutionary timescales can be estimated from DNA sequence data using the molecular clock, a statistical model that describes the behaviour of evolutionary rates among organisms. Although originally based on the assumption of rate constancy among lineages, molecular clocks now include ‘relaxed’ variants that are able to accommodate heterogeneous rates. These have played an important role in resolving evolutionary rates and timescales across the Tree of Life.

Genome-scale data offer exciting opportunities for improving our understanding of molecular evolution and refining our estimates of evolutionary timescales. A large number of genome projects are in the pipeline, promising to produce a flood of data that will be useful for evolutionary inferences. However, they also bring considerable computational and analytical challenges. I describe some of the approaches that have been used to estimate evolutionary timescales from genome-scale data, drawing on several recent examples from my work. I also explain how phylogenetic analysis can provide insight into molecular evolutionary dynamics, with particular reference to the “pacemaker” models of genome evolution.

Biography: Simon is a Professor of Molecular Evolution at the University of Sydney, where leads the Molecular Ecology, Evolution, and Phylogenetics Lab. His research interests include phylogenetic methods, evolutionary models, and molecular clocks. The focus of Simon’s recent work has been on describing patterns of evolutionary rate variation across genomes and on phylogenomic estimation of evolutionary timescales.Simon received his PhD in 2006 (University of Oxford), then did postdoctoral work at Oxford and the Australian National University before joining the University of Sydney in 2010. He has received a number of awards for his work, including the 2015 Edgeworth David Medal, 2014 NSW Young Tall Poppy of the Year, and 2015 Eureka Prize for Outstanding Early Career Researcher.

SPECIAL GUEST PRESENTATION

11:00am – 11.30am, P514

METABOLOMICS – AN IMPORTANT PIECE IN THE 'OMICS PUZZLE

Associated Professor Ute Roessner

University of Melbourne

Abstract: Metabolomics is an emerging field in the suite of 'omics' approaches for Systems Biology. The goal of metabolomics is to detect the presence of all small-molecules in a biological sample. This presents a significant challenge due to their chemical diversity and large concentration ranges requiring the application of complementary analytical approaches, including mass spectrometry coupled to chromatography, to increase the coverage of metabolites analysed. The array of analytical approaches will be summarised and their application in biological systems demonstrated with examples from our research programs. The presentation will focus on computational and statistical challenges metabolomics researchers are facing and will highlight some of the workflows and tools currently available for conventional orthogonal metabolomics analyses. In addition, an outlook on novel techniques for spatial tissue metabolite and lipid analysis will be presented, again with a focus on the computational challenge of comparative highly dimensional mass spectrometry tissue imaging analyses.

Biography: A/Prof Ute Roessner has obtained her PhD in Plant Biochemistry at the Max-Planck-Institute for Molecular Plant Physiology in Germany, where she developed novel GC-MS based methods to analyse metabolites in plants. With the combination of small molecule analytics and sophisticated bioinformatics and statistics the field of metabolomics was born which today is an important tool in biological sciences, systems biology and biomarker discovery. In 2003 she moved to Australia where she established a GC-MS and LC-MS based metabolomics platform as part of the Australian Centre for Plant Functional Genomics (www.acpfg.com.au). Between 2011 and 2014 she led the ACPFG node at the School of BioSciences, The University of Melbourne. Also, in 2007 Ute became involved in the setup of the government funded Metabolomics Australia (MA, www.metabolomics.com.au), a member of BioPlatforms Australia (www.bioplatforms.com.au) and now leads the MA node at the School of BioSciences, The University of Melbourne. In 2013 she was awarded a prestigious ARC Future Fellowship which aims to identify novel mechanisms of salinity tolerance in barley by spatial analysis of metabolites and lipids using Imaging Mass Spectrometry.

SPECIAL GUEST PRESENTATION

4.10pm - 4:50pm, P514

Proudly supported by Bioplatforms Australia

BIOPLATFORMS AUSTRALIA

The Road to Utopia:

Challenges in linking literature and research data

Professor Terri Attwood

Professor of Bioinformatics

University of Manchester

Abstract: This presentation describes a personal journey in which a lifetime preoccupation with protein sequences and databases led to the unlikely development of Utopia Documents, a 'smart' PDF reader (http://getutopia.com), and a pioneering project to create the Semantic Biochemical Journal (http://www.biochemj.org/bj/semantic_faq.htm). Our quest was motivated by a desire to better link data and/or software tools with scientific articles, to blur the boundaries between databases and papers. During an era of ‘big data’, when more articles are being published and more data are being produced than humans can readily assimilate, Utopia Documents offers a new paradigm for extracting nuggets of information from the barrage of published scientific information that now assaults us every day. Examples will be given from the Lazarus project, in which Utopia harnesses the power of the ‘crowd’ to capture asserted facts and relationships automatically, as a simple side-effect of reading and interacting with scientific articles.

Biography: Currently, I’m Professor of Bioinformatics at the University of Manchester. Research interests in protein sequence analysis and protein family classification led to the development of databases like PRINTS & InterPro, and software tools like CINEMA & Utopia; more recently, an interest in the area of semantic integration of research data with scholarly publications led to the creation of Utopia Documents and launch of ‘The Semantic Biochemical Journal’ with Portland Press. When not involved in research, I’m a keen educator and trainer: I wrote the first introductory bioinformatics text-book - my 3rd book was published this summer. In 2012, I led the initiative to create the GOBLET Foundation (the Global Organisation for Bioinformatics Learning, Education and Training), and I currently lead the development of ELIXIR's Training e-Support System, TeSS.

SESSION 5: BRIEF PRESENTATIONS

4 x 15 minute talks (12 mins + 3 mins for questions)

Paper	Presenting author
Investigating the evolution of new biochemical pathways in baker's yeast Saccharomyces cerevisiae	*Åsa Pérez-Bercoff* School of Biotechnology and Biomolecular Sciences, University of New South Wales
Identification of allelic-specific methylation profiles across generations in the Norfolk Island genetic isolate	*Miles Benton* Genomics Research Centre, Queensland University of Technology
Estimating genetic similarity with the k-mer Weighted Inner Product (kWIP)	*Kevin Murray* Centre of Excellence in Plant Energy Biology, Australian National University
De novo characterisation of RNA structure motifs from ENCODE RIPseq data	*Martin Smith* Garvan Institute of Medical Research

ORAL PRESENTATION ABSTRACTS

Investigating the evolution of new biochemical pathways in baker's yeast Saccharomyces cerevisiae

Åsa Pérez-Bercoff (a), Tonia L. Russell (b), Philip J. L. Bell (c), Paul V. Attfield (c) and Richard J. Edwards (a)

(a) School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney NSW, Australia
(b) Ramaciotti Centre for Genomics, University of New South Wales, Sydney NSW, Australia.
(c) Microbiogen Pty Ltd, Sydney NSW, Australia.

Understanding how new biochemical pathways evolve in a sexually reproducing population is a complex and largely unanswered question. We have successfully evolved a novel biochemical pathway in yeast using a sex based population approach.

For over 30 years, wild type Saccharomyces has been widely reported to not grow on xylose at all, but we discovered that most strains can grow, albeit at almost undetectable rates. A mass mated starting population of Saccharomyces cerevisiae strains was evolved under selection on Xylose Minimal Media (XMM) with forced sexual mating every ~two months for 1463 days. This produced a population that could grow on xylose as a sole carbon source. Initial studies show the xylose growth trait is quantitative and presumably governed by many genes. To investigate the evolution of the xylose phenotype, a xylose utilising strain MBG11a was isolated. MBG11a was sequenced with PacBio RSII long read sequencing at the Ramaciotti Centre for Genomics. A high quality complete genome was assembled de novo using the hierarchical genome-assembly process (HGAP3) using only PacBio non-hybrid long-read SMRT sequencing data, corrected using Quiver, and compared to the genome of the S. cerevisiae S288C reference genome.

Approximately 98.5% of the MBG11a genome could be aligned to S288C at 99.5% sequence identity, with over 15,000 non-synonymous and 200 nonsense SNP differences. We have crossed MBG11a with a reference wild type yeast strain (X2180 gal2, Xyl-) and are testing offspring on different minimal media in an attempt to identify MBG11a variants responsible for the novel growth phenotype.

Understanding what has occurred in the evolving yeast population, and how the yeast genome adapted under the selection pressures is of broad interest as it allows experimental analysis of how novel complex biological functions can evolve in an organism.

Identification of allelic-specific methylation profiles across generations in the Norfolk Island genetic isolate

Miles Benton (a), Rod Lea (a), Nicole White (b), Daniel Kennedy (b), Heidi Sutherland (a), Larisa Haupt (a), Kerrie Mengersen (b) and Lyn Griffiths (a)

(a) Genomics Research Centre, Institute of Health and Biomedical Innovation, Queensland University of Technology, Brisbane, Queensland, Australia
(b) ARC Centre of Excellence for Mathematical and Statistical Frontiers, Queensland University of Technology (QUT), Brisbane, Queensland, Australia

DNA methylation is an important epigenetic mechanism that can contribute to variation in gene expression and complex traits. Characterising the genome-wide diversity of DNA methylation and understanding allele-specific inheritance patterns across generations is the next frontier in human genetics. The use of genetic isolates provides an innovative natural setting for investigating DNA methylation because of large multi-generational pedigrees and reduced genetic and environmental diversity. The Norfolk Island (NI) population is one such genetic isolate. Located off the east coast of Australia, the original population was founded by 11 British mutineers of the HMS Bounty and 6 Polynesian women in the late 1700s. These founders have given rise to a 6000-member pedigree spanning 11 generations. We are measuring genome-wide allele-specific methylation using NGS bisulphite sequencing (CpGiant with Illumina HiSeq). To date we have collected data for 24 individuals comprising a close 3 generation pedigree. Data was analysed using a custom pipeline incorporating Methpipe software for calling allele-specific methylation (ASM). Results of this analysis identified 1.12M CpG sites common across all samples. Of these 257992 CpG sites showed ASM. Using a custom clustering method we identified ~1800 ASM regions (AMRs) conserved within the pedigree. Many of these AMRs map to known and predicted imprinted genes. Interestingly, there were numerous ASM peaks in the NI family that have not been previously identified as imprinted loci. For example, we observed a large AMR peak on Chr 21 with no previous imprinting information and very little SNP annotation. This peak also demonstrated 'conservation' of ASM signal across all samples, and looks to be in a 50:50 split - typical of imprinting. This finding provides compelling evidence that our novel approach has potential to identify new imprinted genes as well as allow us to characterise trans-generational epigenetic inheritance patterns that might influence complex traits in humans.

Estimating genetic similarity with the k-mer Weighted Inner Product (kWIP)

Kevin Murray (a), Cheng Soon Ong (b, c), Christfried Webers (b,c), Justin Borevitz (a), Norman Warthman (a)

(a) Centre of Excellence in Plant Energy Biology, The Australian National University, Canberra, Australia
(b) Data61, CSIRO, Canberra, Australia
(c) Department of Computer Science, The Australian National University, Canberra, Australia

Modern genomics techniques generate overwhelming quantities of data. Extracting population genetic variation demands computationally efficient methods to determine genetic relatedness between individuals or samples in an unbiased manner, preferably de novo. The rapid and unbiased estimation of genetic relatedness has the potential to overcome reference genome bias, to detect mix-ups early, and to verify that biological replicates belong to the same genetic lineage before conclusions are drawn using mislabeled, or misidentified samples.

We present the k-mer Weighted Inner Product (kWIP), an assembly-, and alignment-free estimator of genetic similarity. kWIP combines a probabilistic data structure with a novel metric, the weighted inner product (WIP), to efficiently calculate pairwise similarity between sequencing runs from their k-mer counts. It produces a distance matrix, which can then be further analysed and visualized. Our method does not require prior knowledge of the underlying genomes and applications include detecting sample identity and mix-up, non-obvious genomic variation, and population structure.

We show that kWIP can reconstruct the true relatedness between samples from simulated populations. By re-analyzing several published datasets we show that our results are consistent with marker-based analyses. kWIP is written in C++, licensed under the GNU GPL, and is available from https://github.com/kdmurray91/kwip.

De novo characterisation of RNA structure motifs from ENCODE RIPseq data

Martin A. Smith (a), Stefan E. Seemann (b), Luis R. Arriola-Martinez (a), Xiucheng Quek (a), John S. Mattick(a)

(a) Garvan Institute of Medical Research & St-Vincent's Clinical School, Faculty of Medicine, UNSW Australia
(b) Centre for non-coding RNAs in disease and health, Faculty for Health and Medical Sciences, University of Copenhagen, Denmark

Most of the human genome is transcribed into RNA while less than 2% produces protein coding genes. Despite their ever increasing prevalence in reference transcriptomes, relatively few long non-coding RNAs have been functionally characterised to date. Understanding how ncRNAs function is essential to improve our understanding of normal biology and disease, as exemplified by the 4:1 ratio of disease-associated mutations arising in non-coding vs. protein coding regions in the human genome. All ncRNAs form higher-order structures via base-pairing, which span over 13% of the human genome as we have previously shown by measuring the evolutionary hallmarks of structured RNAs in multiple genome alignments. To assign discrete functions to RNA structure predictions, we developed a pipeline for the discovery of recurring motifs in RNA binding protein data. At its core lies DotAligner, a lightweight RNA base-pairing probability matrix alignment heuristic that can classify RFAM sequences with exceptional precision and speed. We applied this pipeline to public ENCODE eCLIPseq (enhanced uv Cross-Linking ImmunoPrecipitation) data for 44 RNA-binding proteins, identifying known and novel clusters of structurally homologous RNA motifs. Furthermore, this agnostic approach reveals RNA structures bound by distinct proteins known to form protein-protein interactions, suggesting that quaternary protein structure is involved in targeting specific nucleotide structures. The resulting clusters of structural motifs can then be used to generate covariance models and scan the genome for homology. Our preliminary results have revealed thousands of new hits, increasing the genomic coverage of the original queries by 150x. We expect that this work will facilitate the attribution of specific biological functions to lncRNAs in a systematic manner.

SESSION 6: BRIEF PRESENTATIONS

4 x 15 minute talks (12 mins + 3 mins for questions)

Paper	Presenting author
EMBL Australia Bioinformatics Resource (EMBL-ABR)	*Pip Griffin* EMBL Australia Bioinformatics Resource
Through the looking glass: using Monocle to visualise methylation array data	*Jovana Maksimovic* Murdoch Childrens Research Institute
Big data from a little proteolysis: Combining multiple omics platforms to identify novel functions of a protease associated with prostate cancer	*Ruth Fuhrman-Luck* Australian Prostate Cancer Research Centre
Geometric map-reduce algorithms for alignment-free sequence comparison	*Lawrence Buckingham* School of Electrical Engineering and Computer Science, Queensland University of Technology

ORAL PRESENTATION ABSTRACTS

EMBL Australia Bioinformatics Resource (EMBL-ABR)

Pip Griffin (a), Sonika Tyagi (b), Ira Cooke (c), Philipp E. Bayer (d), David Edwards (d), Dominique Gorse (e), Saravanan Dayalan (f), Sylvain Foret (g), Jac Charlesworth (h), Steven Androulakis (i), Marc Wilkins (j), Rob Cook (e), Malcolm McConville (f)

(a) EMBL Australia Bioinformatics Resource: VLSCI Node; (b) EMBL-ABR: AGRF Node; (c) EMBL-ABR: JCU Node; (d) EMBL-ABR: UWA Node; (e) EMBL-ABR:QCIF Node; (f) EMBL-ABR: MA Node; (g) EMBL-ABR: ANU Node; (h) EMBL-ABR: UTas Node; (i) EMBL-ABR: Monash Node; (j) EMBL-ABR: SBI Node; (k) EMBL-ABR Hub

This talk will provide an overview of the recently established EMBL Australia Bioinformatics Resource (EMBL-ABR) and how it aligns with national and pan-national efforts around the globe, such as deNBI, DTL, SIB, ELIXIR, BD2K, CyVerse and Canada B/CB. EMBL-ABR is a distributed national research infrastructure with the mission to provide bioinformatics support to life science and medical researchers in Australia. It is currently composed of ten nodes (expertise centres) and one coordinating HUB hosted by EMBL-ABR: VLSCI node in Melbourne. This talk will also provide an overview of the activities across EMBL-ABR Key Areas: Data, Tools, Platforms, Compute, Training and Standards.

Through the looking glass: using Monocle to visualise methylation array data

Jovana Maksimovic (a), Yuxia Zhang (c,d,e), David Martino (a,c,f,g), Richard Saffrey (a,c), Len Harrison (c,d), Alicia Oshlack (a,b)

(a) Murdoch Childrens Research Institute, Royal Children's Hospital, Parkville, Victoria, Australia
(b) School of BioSciences, The University of Melbourne, Parkville, Victoria, Australia
(c) The University of Melbourne, Victoria, Australia
(d) Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria, Australia
(e) Guangzhou Institute of Pediatrics, Guangzhou Women and Children's Medical Center, Guangzhou Medical University, Guangzhou, China
(f) Centre for Food and Allergy Research, Murdoch Childrens Research Institute
(g) Member of 'In-FLAME' the International Inflammation Network, World Universities Network (WUN)

Bioinformatics software is generally built to analyse or process a specific data type. This is because different data types have different properties that need to be taken into account and many types of analyses are only relevant in a particular data context. However, we recently discovered that it is occasionally beneficial to stray from this established paradigm to generate a more biologically meaningful result from a complex dataset.

The dataset in question is a methylation array dataset (n=195) interrogating sorted immune cells, from 10 different individuals, which have been stimulated with various combinations of cytokines. The cytokine combinations activate cells towards differentiation down several key pathways. Given the large number of samples, standard dimensionality reduction methods such as multi-dimensional scaling (MDS) managed to reveal the gross structure in the data but they lacked the resolution to display the subtle nuances of the cellular differentiation trajectories.

Here we show that using the Monocle method on this dataset reveals important structure unobservable with standard clustering methods. Monocle was originally developed for single cell RNA-seq data for the analysis of "dynamic biological processes such as cell differentiation". As our dataset describes the methylation changes associated with immune cell activation and differentiation, analysing it with Monocle seemed to make biological sense even though the data type was atypical. We transformed the methylation data by removing inter-individual variation and imported it into Monocle. We then used the 1000 most variable CpG loci to order the samples based on their progress through the activation/differentiation process and plot a spanning tree. In contrast to standard methods, the resulting analysis crystallised the biological relationship between the various immune cell types and activation/differentiation conditions.

Big data from a little proteolysis: Combining multiple omics platforms to identify novel functions of a protease associated with prostate cancer

Ruth Fuhrman-Luck (a), Marcus Hastie (b), Thomas Stoll (b), Oded Kleifeld (c), Bosco Ho (d), Melanie Lehman (a), Anja Rockstroh (a), Thomas Kryza (a), Carson Stephens (a), Colleen Nelson (a), Jeffrey Gorman (b), Daniela Loessner (e) and Judith Clements (a)

(a) Australian Prostate Cancer Research Centre – QLD, Translational Research Institute, Queensland University of Technology, Brisbane, Australia
(b) QIMR Berghofer Medical Research Institute, Brisbane, Australia
(c) Faculty of Medicine, Nursing and Health Sciences, Monash University, Clayton, Australia
(d) Monash Bioinformatics Platform, Monash University, Clayton, Australia
(e) Institute of Health and Biomedical Innovation, Queensland University of Technology, Brisbane, Australia

A single protease can irreversibly process a large number of functionally diverse proteins, with far-reaching downstream effects to cell signalling, regulation and function. As such, omics approaches are integral for the study of protease function, for their ability to identify a multitude of protease substrates (degradomics) within a complex protein background, and allow for large-scale analyses of downstream regulatory responses at the gene (transcriptomics) and protein (proteomics) level. We employed multiple omics techniques to analyse the function of kallikrein 4, a serine protease that is up-regulated in prostate cancer and purported to have a functional role in cancer progression.

Two degradomics approaches (TAILS, PROTOMAP) were employed to identify kallikrein 4 substrates produced by prostate cancer cells and prostate myofibroblasts. Transcriptomics analyses (custom Agilent 180K RNA microarray) were also performed to identify kallikrein 4-regulated genes, and kallikrein 4-mediated regulation was confirmed at the protein level by SILAC-proteomics. Finally, bioinformatics tools (IPA, DAVID, TopFIND) were used to develop data-driven hypotheses, which were tested using targeted biochemical and cellular assays.

Over 100 novel putative kallikrein 4 substrates were identified. Particularly, kallikrein 4 activated another prostate cancer-associated protease, matrix metalloproteinase-1, and cleaved protease regulators, growth factors and ECM proteins, among others. Kallikrein 4 also induced extensive gene expression changes in prostate myofibroblasts, and informatics suggested the activation of two signalling pathways implicated in the conversion from a normal fibroblast phenotype to that observed in cancerous tissue. The ability for kallikrein 4 to activate these signalling pathways was confirmed in vitro. Thus, multi-omics platforms were used to provide data-informed hypotheses that were confirmed in vitro, identifying novel mechanisms by which kallikrein 4 may promote prostate cancer progression.

Geometric map-reduce algorithms for alignment-free sequence comparison

Lawrence Buckingham (a), Timothy Chappell (a), Shlomo Geva (a), Paul Greenfield (b), James M. Hogan (a), Wayne Kelly (a), Dimitri Perrin (a)

(a) School of Electrical Engineering and Computer Science, Queensland University of Technology
(b) CSIRO

Sequence comparison is a central component of computational biology, and while alignment-based approaches such as BLAST dominate the field at present, these algorithms do not scale well in the face of the explosion of genomic data which is being generated by current sequencing technologies. Alignment-free algorithms, including several that are derived from refinements of the well-known D_2 statistic, show promise in terms of efficiency and ability to deal with structural rearrangements. These algorithms rely on normalised similarity between word count vectors so they fail when confronted by highly diverged sequences due to dependence on exact word matches.

In the present work we reformulate sequence comparison as a nearest neighbour search between two sets of points in a high-dimensional non-metric space. Similarity measurement is inspired by the well-known Hausdorff set distance. Inexact word matches are handled by means of substitution matrices, enabling use of much longer words than is considered normal in alignment-free methods. Given a pair of sequences, their similarity is calculated by forming the pair-wise word similarity matrix, which then passes through a sequence of reduction phases. Row and column similarity profiles are obtained by taking row- and column-wise maxima. Each profile vector is then subject to a further reduction, either average or maximum. The resulting values are combined by a reduction such as minimum, average, or maximum, to obtain the final result.

We provide theoretical results and demonstrate that the precision of the new algorithm exceeds that of BLAST on a protein homology search task. We also present approximate versions of the algorithm based on sampling and sequence partitioning which offer promise for improved computational efficiency.

SESSION 7: BRIEF PRESENTATIONS

5 x 15 minute talks (12 mins + 3 mins for questions)

Paper	Presenting author
TBA	*COMBINE Student*
GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly	*Daniel L Cameron* Walter and Eliza Hall Institute of Medical Research
Detecting pathogenic STR expansions in next-gen sequencing data	*Harriet Dashnow* Murdoch Childrens Research Institute
BootNet: a bootstrapping application for GLMnet modelling to identify robust classifiers in genomics data	*Rod Lea* Genomics Research Centre, Queensland University of Technology
Assessing the Practicality of Oxford Nanopore Sequencing in Clinical Diagnositcs	*Alexis Lucattini* Australian Genome Research Facility

ORAL PRESENTATION ABSTRACTS

Selected presentation from Monday's COMBINE 2016 Symposium

Presenter to be announced

Affiliations to be announced

COMBINE 2016 fields an exciting and diverse array of presentations from students and early career researchers. One of these presenters will be chosen to give a repeat performance at AB^{3}ACBS.

GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly

Daniel L Cameron (a,b), Anthony T Papenfuss (a,b,c)

(a) Bioinformatics Division, Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria, 3052, Australia
(b) Department of Medical Biology, University of Melbourne, Parkville, Victoria, 3010, Australia
(c) Peter MacCallum Cancer Centre, Victorian Comprehensive Cancer Centre, Melbourne, 3000, Australia

The identification of genomic rearrangements with high sensitivity and specificity using massively parallel sequencing remains a major challenge. Many methods have been developed for Illumina sequence data, with most methods using read depth analysis, read pair clustering, split read identification, assembly, or a combination of these approaches. Existing assembly-based methods perform either de novo assembly (e.g. cortex), targeted assembly based on previously identified candidates (e.g. manta, SVMerge, TIGRA), or perform windowed assembly to detect small events (e.g. DISCOVAR, SOAPindel).

Here we describe GRIDSS, the Genome Rearrangement IDentification Software Suite, composed of an assembler, and a variant caller which combines assembly, split read and read pair evidence to identify genomic rearrangement breakpoints using a probabilistic model. Our novel genome-wide break-end assembly approach assembles reads not supporting the reference prior to breakpoint identification or variant calling using a positional de Bruijn graph. By constraining the assembly of each read based on the mapping locations of the read/read pair, and encoding these assembly constraints directly within the assembly graph itself, a single genome-wide assembly can be performed.

GRIDSS recently won structural variant detection sub-challenge #5 of the ICGA-TCGA DREAM Somatic Mutation Calling challenge and has been extensively benchmarked against BreakDancer, cortex, CREST, DELLY, HYDRA, LUMPY, manta, Pindel, SOCRATES and TIGRA across a wide range of simulated variant types, variant sizes, read depths, read lengths, and library fragment sizes. With the exceptions of low coverage data (≤8x), and large novel insertions detectable only by de novo assemblers, GRIDSS F-scores exceeded that of all other callers. On well-studied human cell line data, GRIDSS is able to achieve a false discovery rate less than half that of other methods, with no loss of sensitivity.

Detecting pathogenic STR expansions in next-gen sequencing data

Harriet Dashnow (a,b) and Alicia Oshlack (a,b)

(a) Murdoch Childrens Research Institute, Royal Children's Hospital, Victoria, Australia
(b) Biosciences, University of Melbourne, Parkville, Victoria, Australia

Short tandem repeats (STRs) are 1-6bp DNA sequences repeated in tandem. STR expansions are known to cause more than 25 Mendelian diseases, most notably Huntington's disease, spinocerebellar ataxias, and the fragile X disorders. Current methods for detecting pathogenic STR expansions involve PCR amplification and electrophoresis, and a specific assay must be designed for each locus. These methods do not scale to the whole genome and so cannot be used to identify new pathogenic STR loci.

As exome and whole genome sequencing is becoming increasingly common in the search for the genetic causes of human disease, we need methods capable of detecting pathogenic STR expansions at the genome-wide scale. There are a number of tools for genotyping STR variation in such data – most notably LobSTR and RepeatSeq – however these tools can only detect STR variants shorter than the read length. However most pathogenic variants involve a significant increase in length, far exceeding current Illumina read lengths. Tools to genotype STRs longer than the read length still require the variant to be within the insert size, and require tight insert size distributions, which are becoming increasingly rare in recent Illumina protocols.

Here we describe a method to detect pathogenic STR expansions at all known STR loci in the genome from next generation sequencing data. This method uses decoy sequences added to the reference genome to identify reads originating from a large STR expansion, then identifies their source using paired information. We have validated this method on simulated exome data and applied it to assess the STR variation in an autism cohort.

BootNet: a bootstrapping application for GLMnet modelling to identify robust classifiers in genomics data

Rod Lea (a), Nicole White (b), Ray Blick (c), Macartney-Coxson (c), Daniel Kennedy (b), Lyn Griffiths (a), and Miles Benton (a)

(a) Genomics Research Centre, Institute of Health and Biomedical Innovation, Queensland University of Technology, Brisbane, Queensland, Australia
(b) ARC Centre of Excellence for Mathematical and Statistical Frontiers, Queensland University of Technology (QUT), Brisbane, Queensland, Australia
(c) Kenepuru Science Centre, Institute of Environmental Science and Research, Wellington, New Zealand

The R-based GLMnet package is a powerful machine learning classifier that provides benefits to analysis of highly dimensional genomic data through the implementation of an elastic-net routine. This framework brings together two established approaches, penalised ridge regression and LASSO, and by applying specific tuning parameters is able to overcome limitations conventional methods such as OLS regression. We developed a wrapper to GLMNet (BootNet) that adds fast bootstrap resampling algorithms to all types of outcome and predictor variables. BootNet accepts both quantitative and qualitative outcomes, i.e. tissue type, disease state, blood pressure, age. The base algorithm employs a percentage-based sub-sampling routine. We have also implemented a Jackknife approach (for outlier detection) as well as a leave-one-out cross-validation method. Additionally, we have added the ability to run processing in parallel, greatly reducing the time of computation on both local machines and servers. In testing we have explored numerous methylation data sets from various tissues with great success. Using a discovery cohort we were able to select CpG sites that differentiate abdominal from omental adipose 100% of the time. In another experiment we demonstrated that the BootNet method could identify robust DMRs and single CpGs associated with aging in a healthy cohort, with potential biological relevance and statistical significance. We were able to validate our top findings in the large cohort (n=2316) of publicly available methylation data from MARMAL-AID. The elastic-net methods implemented in GLMnet are very powerful for classifying data, and with our BootNet wrapper this package now has potential to be developed into a fast integrative 'omics approach for identification of robust classifiers of states such as disease outcomes.

Assessing the Practicality of Oxford Nanopore Sequencing in Clinical Diagnositcs

Alexis Lucattini (a,b), Lavinia Gordon (a), Matt Ritchie (b)

(a) Australian Genome Research Facility
(b) Walter & Eliza Hall Institute for Medial Research

Much excitement has arisen over the new Oxford Nanopore's MinION and it's potential in real-time diagnostic sequencing applications. Unfortunately in order to obtain high-molecular DNA many samples provide insufficient DNA to extract and sequence directly, requiring overnight growth following by extraction. Thus hindering the real-time advantage of the platform. The new Oxford Nanopore R9 kit comes with the potential of 'rapid sequencing', requiring only a ten minute library-preparation. Using the QIAGEN-Mini Kit we were able to obtain a sufficient DNA extraction within 120 minutes followed by the rapid 1D preparation. With local base calling, we compared the output of using a local metagenomic database; OneCodex; and WIMP (What's in my pot?) Oxford Nanopore's metagenomic classifier. We found that the OneCodex database is the most extensive of the three, with a much more sensitive aligner. Our own local database allowed for fast identification but only classified 5% of the reads. WIMP had a similar classification level to our own local database but lagged due to upload/download time. We present all three using Krona, which displays interactive hierarchical data on zoomable piecharts.

SESSION 8: BRIEF PRESENTATIONS

3 x 15 minute talks (12 mins + 3 mins for questions)

Paper	Presenting author
How do vulnerabilities left during the evolution of cellular networks set the stage for cancer?	*Anna Trigos* Bioinformatics and Cancer Genomics Laboratory, Peter MacCallum Cancer Centre
TCP-seq, a novel technique for investigating mechanisms and regulation of eukaryotic translation initiation	*Stuart Archer* The John Curtin School of Medical Research, Australian National University

ORAL PRESENTATION ABSTRACTS

How do vulnerabilities left during the evolution of cellular networks set the stage for cancer?

Trigos, A.S. (a,b), Pearson, R.B. (b,c,d), Papenfuss, A.T. (a,b,e), Goode, D.L. (a,b)

(a) Bioinformatics and Cancer Genomics Laboratory, Peter MacCallum Cancer Centre, Melbourne, Victoria 3000, Australia.
(b) Sir Peter MacCallum Department of Oncology, The University of Melbourne, Victoria 3010, Australia.
(c) Department of Biochemistry and Molecular Biology, The University of Melbourne, Parkville, Victoria 3010, Australia.
(d) Department of Biochemistry and Molecular Biology, Monash University, Clayton, Victoria 3168, Australia.
(e) Bioinformatics Division, The Walter & Eliza Hall Institute of Medical Research, Parkville, Victoria 3052, Australia.

Taking a top-down approach from protein-protein and gene-regulatory networks to derive biologically relevant information and key genes is challenging given their complexity. To bridge the gap between network-level and gene-level analysis, we tackled an intermediary level of organization, interactions between biological processes.

We developed a metric for the interconnectedness of high-level cellular processes (GOslims) based on the entire human network of gene and protein interactions to construct a network of co-regulation. This metric normalizes the number of gene-gene edges connecting pairs of processes by the total possible number of edges. Overlaying co-expression data from 7 tissue types onto our co-regulation network identified patterns of transcriptional and functional dependency.

A strong positive correlation of expression of highly interconnected processes involved in basal cellular functions suggests co-regulation developed by a long history of co-evolution. In contrast, consistent negative correlation in expression indicates incompatibility between the activation of basal and more recently evolved cellular functions, suggesting strong mutual exclusivity.

Expression data from The Cancer Genome Atlas demonstrated how these patterns are altered in cancer. We detected an enhanced mutual exclusivity of unicellular and multicellular functions, with many pairs of processes key to carcinogenesis positively correlated in normals becoming negatively correlated in tumours. Extending our approach to the gene level, we found the same genes modulated these alterations across multiple tumours, uncovering putative novel common drivers.

Our results indicate the co-regulatory networks between cellular processes can modulate adaptive and evolutionary processes, and their disruption can drive tumour progression. The study of these networks provides a level of organization that facilitates the link between network-level and gene-level analyses.

TCP-seq, a novel technique for investigating mechanisms and regulation of eukaryotic translation initiation

Stuart K. Archer(a,b), Nikolay E. Shirokikh(a,c), Steve Androulakis(b), Traude H. Beilharz(d), and Thomas Preiss(a,e)

(a) EMBL–Australia Collaborating Group, Department of Genome Sciences, The John Curtin School of Medical Research, The Australian National University, Canberra, Australia
(b) Monash Bioinformatics Platform, Monash University, Melbourne, Victoria 3800, Australia.
(c) Moscow Regional State Institute of Humanities and Social Studies, Kolomna 140410, Russia.
(d) Development and Stem Cells Program, Monash Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology, Monash University, Melbourne, Victoria 3800, Australia.
(e) Victor Chang Cardiac Research Institute, Darlinghurst, New South Wales 2010, Australia.

Eukaryotic translation initiation is a fundamental part of the central dogma of molecular biology, and an important point of gene expression regulation, however there are large blind spots in our knowledge of both the basic mechanism and the points in the process at which regulatory intervention can influence gene expression. We describe the analysis of the first data from TCP-seq, a novel variant of RNA-seq that gives a transcriptome-wide snapshot of translation initiation events. By analysing the position and length of RNase-protected 'footprints' left behind by the small ribosomal subunit, which is the first ribosomal subunit to bind mRNA, general features of initiation and start-codon recognition can be discerned. Transcript-specific departures from the normal initiation processes are also apparent in many cases. Valuable information about the states of the original complexes can be inferred from the lengths of the RNA footprints, unlike in conventional RNA-seq which is primarily concerned with fragment location only. Therefore, novel analysis pipelines were developed for both the metagene and the gene-by-gene analyses to draw correlations between lengths and positions of RNA footprints relative to the start codon. We thus inferred conformational changes in the initiation complex as it proceeds in the 5' to 3' direction, scanning for the start codon. A Shiny app was also developed to visualize the data on a gene-by-gene basis. The results underpin mechanistic models of translation initiation and termination, built on decades of biochemical and structural investigation, with direct in vivo evidence on a transcriptome-wide scale.

POSTER PRESENTATION ABSTRACTS

Poster ID: 1 (COMBINE)

Prediction of disulfide bond dihedral angles based on chemical shifts

Quentin Kaas, Johan Rosengren

(a)School of Biomedical Sciences, The University of Queensland, Brisbane, Queensland, 4072, Australia
(b)Institute of Molecular Biosciences, The University of Queensland, Brisbane, Queensland, 4072, Australia
(c)School of Biomedical Sciences, The University of Queensland, Brisbane, Queensland, 4072, Australia

Peptide toxins are attractive drug leads due to their high selectivity and potency for receptor targets. They often contain multiple disulfide bonds, resulting in highly restrained structures. Despite their potential, only a limited number of toxin analogues have made it to market because of limitations in terms of stability and delivery. Therefore, there is an obvious need for continual research into the understanding of structure activity relationships and methods for developing improved analogues. An essential requirement for development of drug leads is the elucidation of a complete and precise 3D structure of bioactive peptides. A well-proven technique for structural elucidation is that of NMR spectroscopy. Despite being a major structural determinant, the restraints imposed on disulfide bonds during structural calculations are currently limited. Based on a proposed correlation between disulfide bond configuration and chemical shifts, we developed a novel support vector machine that was able to successfully predict all five χ dihedral angles. We showed how these restraints can reduce ambiguity and increase the precision of final peptide structures, underlining a method that can improve the rational drug design process.

Poster ID: 2 (COMBINE)

Identification of the dominant endogenous factors regulating inflammation and regeneration in skeletal muscle following physical trauma

Lian Liu (a), Jonathan Peake (a) and Tony Parker (a)

Tissue Repair and Regeneration Program, Institute of Health and Biomedical Innovation, School of Biomedical Sciences, Faculty of Health, Queensland University of Technology, Brisbane, QLD

The project will combine global/profiling and functional proteomics/metabolomics based approaches in in vivo model. Specifically, a rat impact contusion model will be utilised to model impact trauma. Dynamic global protein and metabolite profiling using, LC-MS/MS and NMR respectively, will be performed at 6h, 12h, 1, 2, 3, 7 and 14 days on tissue homogenates to identify factors that are associated with the initial recovery response following injury.

Expected Results

1. There are temporal changes to biochemical pathways and processes during recovery from muscle trauma in vivo, and global proteomics/metabolomics profiling will reveal these changes.

2. Dominant endogenous factors present in injured muscle will activate signalling pathways resulting in secretion of cytokines that regulate inflammation and regeneration in skeletal muscle.

Conclusions

This project will use a multi-disciplinary approach, coupled with in vivo experimental model to determine which proteins and/or biochemical pathways facilitate muscle cell functional responses during and following traumatic impact injury.

Poster ID: 3 (COMBINE)

Uncovering Host-Pathogen Interactions in Severe Malaria Through Dual-RNA Sequencing

Hyun Jae Lee (a), Athina Georgiadou (b), Lachlan Coin (a), Michael Levin (b), Thomas Otto (c,d), Michael Walther (e), David Conway (d), Aubrey Cunnington (b)

(a) Institute for Molecular Bioscience
(b) Section of Paediatrics, Imperial College London
(c) Wellcome Trust Sanger Centre
(d) London School of Hygiene and Tropical Medicine
(e) MRC Gambia Unit

The pathogenesis of severe malaria (SM) is incompletely understood. Dual-RNA sequencing can identify differences in the simultaneous host and parasite gene expression profiles of infected individuals to help unravel interactions leading to SM. Presentation blood samples from 46 Gambian children with Plasmodium falciparum malaria (25 SM; 21 uncomplicated (UM)) yielded on average 36 million reads (~70% human and ~25% parasite). Using reference expression signatures for human leukocytes and parasite developmental stages we established that inter-individual heterogeneity in leukocyte and parasite populations explained a large proportion of the variation in gene expression between samples. After adjustment for developmental stage we identified 239 parasite genes differentially expressed between SM and UM, with notable upregulation in RNA processing and downregulation in pyruvate metabolism. After adjustment for leukocyte proportions 771 human genes were differentially expressed. Here SM was associated with genes controlling RNA translation, signatures of innate and B-cell activation, but reduced T cell signaling. Co-expression analysis revealed clusters of strongly correlated host and parasite genes, which indicate that specific interactions may drive different components of pathogenesis. Biological validation is underway.

Poster ID: 4 (COMBINE)

10 things no one tells you about your PhD, but you should know

Paula Andrea Martinez

The University of Queensland

People tell you often to not procrastinate, but anyway you do. People tell you to write now, and still you leave it for tomorrow. They say it is a journey, and it is different in many aspects for every person and for every circumstance, agreed. So many people tell you several tips that you already know about your PhD, but leave unsaid important ones. Point number ten (not to reveal all of them) is “You will feel the most satisfaction when you solved someone’s question/problem by using something you have learned or produced from your PhD”. Whatever your topic is, you are either trying to understand a question or solve a problem, if during your candidature you solved one real life question for someone to act upon this information you have given back to the world. We struggle, most of us, specially towards the end of the PhD candidature. But, hey! we can make it! With this list of 10 things the aim is to help chin up through those rough PhD times.

Poster ID: 5 (COMBINE)

Bioinformatic investigation of the influence of single nucleotide polymorphisms on the phenotype of venous leg ulcers and their healing trajectory

Elizabeth Sydes (a, c), Daniel Broszczak (a), Dianne Maresco-Pennisi (b), Christina Parker (b), Tony Parker (a)

(a) Tissue Repair & Regeneration Program, Injury Prevention & Trauma Management Theme, Institute of Health & Biomedical Innovation, Queensland University of Technology, Brisbane, Queensland, Australia
(b) School of Nursing, Faculty of Health, Queensland University of Technology
(c) Wound Management Innovation Cooperative Research Centre, West End, Queensland, Australia

Venous leg ulcers (VLU) are debilitating wounds that can remain unhealed for several decades and recur in up to 70% of cases. The progression and recalcitrance of the condition is not well understood and, although there is evidence to suggest a genetic predisposition, the genes involved have yet to be elucidated. In this study, a suite of six candidate genes has been selected based on their proposed involvement with venous health. Using qPCR, these genes will be analyzed in VLU patients and age matched controls. Patient wound fluid samples will be analyzed using quantitative liquid chromatography tandem mass spectrometry (LC-MS/MS) to produce protein profiles of approximately 80 ulcers. In particular, the abundance of proteins associated with the target genes and related biochemical pathways will be a focus of this study in addition to comparison of wound fluid biochemistry to VLU associated clinical parameters. Specifically, bioinformatic integration of genetic and proteomic datasets will be performed to determine if there is a genetic link to the biochemistry that underpins non-healing venous leg ulcers. This will be performed using gene ontology and pathway enrichment in applications such as Reactome, Ingenuity Pathway Analysis and Cytoscape. We hypothesize that the genomic analysis will demonstrate a differential expression of wildtype and variant genes between the control and patient cohorts for some or all of the six target genes. Moreover, it is expected that quantitative proteomics of ulcer fluid will show differential abundance in the protein profile between healing ulcers and recalcitrant wounds. In addition, it is anticipated that proteins associated with the target genes will have an altered abundance in non-healing or slow healing wounds. Finally, through integration of our datasets, we expect to find that patients with variant copies of genes will display clinical symptoms that correlate with the physiological role or pathway of the gene as measured at the protein level.

Poster ID: 6 (COMBINE)

Genetic linkage mapping in outcrossed polyploid populations using Genotyping-By-Sequencing and assembly graphs

Chenxi Zhou (a), Wolfgang Gruneberg (b), Federico Diaz (b), Maria David (b), Awais Khan (b), Lachlan JM Coin (a)

(a) Institute for Molecular Bioscience, University of Queensland, Brisbane, Australia
(b) International Potato Center (CIP), Apartado 1558, Lima 12, Peru

Genetic mapping for plant genomes remains challenging due to polyploidy and difficulty in generating inbred mapping populations. Here we describe an approach for constructing genetic linkage maps for polyploid species using genotyping-by-sequencing (GBS) data from outcrossing mapping populations coupled with high coverage whole genome sequence data of a reference genome. Our approach combines de novo assembly using assembly graphs with linkage mapping to arrange scaffolds into pseudomolecules.

By mapping GBS reads to scaffolds we are able to infer scaffold haplotypes even in the presence of substantial amounts of missing data. We use these haplotypes to infer linkage groups corresponding to chromosomes as well as the optimal ordering of scaffolds within these chromosomes.

We show that the method is able to reconstruct simulated chromosomes for both diploid and tetraploid genomes. We compared our method with three existing genetic mapping tools designed for outcrossing species. The results show that our method outperforms the other methods in accuracy on both grouping and ordering. We applied our method to two real datasets - a diploid I. trifida and a tetraploid potato mapping population. The linkage maps show significant concordance with the actual chromosomes. Moreover, we detected and resolved misscaffolding for the de novo genome assembly. We anchored an unplaced scaffold for the PGSC v4.03 pseudomolecules.

Poster ID: 7 (COMBINE)

Chromosome end extension revealed by analysis of completed chromosome end sequences

Haojing Shao, Chenxi Zhou, Lachlan Coin

Institute for Molecular Bioscience, The University of Queensland

The reference human genome is almost complete after many years of effort. However, the chromosome end region is not all fully sequenced possibly due to the divergence of chromosome end and difficulty to assembly long repetitive regions. Here, we use recently well assembled bionano data (N50:3.3MB) in eight samples to anchor and extend the chromosome end. We found that 63.5\% of chromosome end are different from reference with deletion or extension. The deletion and extension sequence are divergence and inheritable inside trios. The chromosome end diversity in 1000 genome declines sharply towards telomere, suggesting the chromosome end are relatively extension sequence in the genome. Analyzing 11 mammals genome chromosome end confirmed the extension. The human extension rate since Hominidae is estimated at 78.30 to 436.1 kbps per million years. The extension could have a functional role by creating new fusion genes. To fit all the observations, we proposed a mechanistic model, namely extra-chromosomal addition.

Poster ID: 8 (COMBINE)

Identification of phylogenetically useful loci

Bokyung Choi (a), Mike Crisp (a), Lyn Cook (b), Robert Edwards (c), Alicia Toon (b), Carsten Külheim (a)

(a) Division of Evolution, Ecology and Genetics, Research School of Biology, The Australian National University, Canberra, ACT, 0200, Australia
(b) School of Biological Sciences, The University of Queensland, Brisbane, QLD, 4072, Australia
(c) Smithsonian Institution, National Museum of Natural History, Washington D.C., the United States.

Sanger sequencing allows researchers to sequence one gene region at a time. However, with the introduction of Next Generation Sequencing it is now possible to sequence multiple gene regions at a time more affordably compared to the traditional sequencing method. Using a small number of markers researchers often could not resolve phylogenetic relationships especially if the gene regions do not have enough phylogenetic signals. This was also the case in our target plant genera Melaleuca and Eucalyptus.

In this study, using the Next Generation Sequencing (NGS) we aimed to identify and sequence orthologous and low copy nuclear exonic gene regions to better resolve phylogenetic relationships within and across the groups. To identify nuclear gene regions that are informative for phylogenetic analysis, we used 1) the assembled transcriptome of M. quinquenrvia as a reference sequence, 2) shot gun sequences of M. bracteata and M. leucadendra, 3) the annotated genome of Eucalyptus grandis and plant genome resources from phytozome (www.phytozome.com). We also introduce how we identified chloroplast gene regions to target by aligning three chloroplast genomes from two Melaleuca species and Eucalyptus grandis. We successfully identified and sequenced 209 gene regions across the groups. This method is potentially useful for researchers who are interested in finding hundreds of loci to resolve phylogenies in plants or other groups of organisms. As more genomic resources (transcriptomes, annotated genomes) are becoming freely available online the method can be even more useful.

Poster ID: 1 (AB3ACBS)

Computational workflows for research students: towards a reproducible research

Soroor Hediyeh-Zadeh and Melissa J Davis

The Walter and Eliza Hall Institute of Medical Research

The majority of new bioinformatics research students are often not acquainted with the common computational workflows and resources. Students may spend several months writing scripts to perform routine analysis, for which a comprehensive set of tools might have already been developed.

Bioconductor is an open source Bioinformatics software (in R language) repository, which provides tools and resources for almost any type of Bioinformatics data analysis, from sequence analysis and differential gene expression to proteomics and metabolomics data analysis. As a very simple example, Bioconductor provides data structures to store gene expression data, the associated gene annotation and the sample information all in one R object (e.g. ExpressionSet object for microarray data, and summarizedExperiment object for RNA-Seq/count data). This ensures consistency when subsetting is done on one table. In addition to tools for reading and pre-processing different data formats (e.g. CELL, SAM/BAM, GTF/GFF, FASTA/FASTQ etc.) many of the command-line tools are now available through Bioconductor, and can be accessed directly from R. It is also possible to retrieve the genomic sequence/location and annotation data for a vast number of species and genome builds. Other public datasets provided by UCSC, NCBI, Refseq etc. can also be accessed directly via Bioconductor packages. In addition, GEO datasets can be downloaded and accessed directly from an R session. All these different tools minimize manual file handling and ensure a reproducible research.

In this talk/poster, the core Bioconductor technologies are introduced. The term Reproducible Research is defined. I, then, introduce Docker and explain how it can be deployed to build a digital archive that reproduces the research results. Finally, the advantages and disadvantages of using Docker over other tools for research reproducibility are discussed.

Poster ID: 2 (AB3ACBS)

Applying machine learning to GWAS analysis of major memory traits

Nesli Avgan (a), Rodney A Le (a), Miles Benton (a), Heidi G Sutherland (a), Lauren G Spriggens (b) David HK Shum (b), Larisa M Haupt (a), Lyn R Griffiths (a)

(a) Genomics Research Centre, Chronic Disease and Ageing, Institute of Health and Biomedical Innovation, School of Biomedical Sciences, Queensland University of Technology, Brisbane, Australia (b) Behavioral Basis of Health Program, Menzies Health Institute Queensland, Griffith University, Gold Coast, Australia

Genome wide association studies (GWAS) is a well-established design for identifying genetic variants that are associated with complex common disease traits. However, there are statistical limitations in GWAS studies, such as multiple testing burden and reduction of power, detecting the effects of SNPs separately and small effect size of the detected SNPs and integrating gene-gene and gene-environmental interactions. In an effort to overcome these limitations we are applying a machine learning approach called GLMNet. Briefly, GLMNet offers extremely efficient procedures for fitting the entire lasso or elastic-net regularization path for regression models. We will apply GLMNet to SNP data for a GWAS study of 619 healthy individuals measured for a battery of tests for evaluating twenty different memory performance phenotypes including visual, episodic and prospective memory. We will compare the GLMNet results to our results of a conventional analysis, performed using PLINK. These results showed highly significant SNP associations for several memory traits but failed to explain a high proportion of genetic heritability of these traits. Here, our goal is to use GLMNet to identify SNP signatures that are highly predictive of these complex memory traits and in turn improve the biological understanding of human memory.

Poster ID: 3 (AB3ACBS)

A transcriptional signature for TGFβ-induced epithelial-mesenchymal transition in cancer

Momeneh Foroutan (a,b), Joseph Cursons (b,c,d), Soroor Hediyeh-Zadeh (b), Erik W. Thompson (a,e,f), Melissa J. Davis (b,g)

(a) The University of Melbourne Department of Surgery, St. Vincent's Hospital, Parkville, VIC 3010, AUSTRALIA; (b) Division of Bioinformatics, Walter and Eliza Hall Institute of Medical Research, Parkville, VIC 3010, AUSTRALIA; (c) Systems Biology Laboratory, Melbourne School of Engineering, The University of Melbourne, Parkville, VIC 3010, AUSTRALIA. (d) ARC Centre of Excellence in Convergent Bio-Nano Science and Technology, Melbourne School of Engineering, The University of Melbourne, Parkville, VIC 3010, AUSTRALIA. (e) Institute of Health and Biomedical Innovation and School of Biomedical Sciences, Queensland University of Technology, Kelvin Grove, QLD 4059, AUSTRALIA (f) Translational Research Institute, Wooloongabba, QLD, 4102, AUSTRALIA (g) Department of Biochemistry and Molecular Biology, Faculty of Medicine, Dentistry and Health, University of Melbourne, Parkville, VIC 3010, AUSTRALIA.

Epithelial-mesenchymal transition (EMT) is a developmental program subverted by cancer cells as they progress through the metastatic cascade. EMT is associated with an aggressive, migratory cell phenotype and it can contribute to clinical chemotherapeutic resistance and poor treatment outcomes. EMT can be induced by different stimuli, of which transforming growth factor β (TGFβ) is one of the best studied. We have used meta-analysis methods to identify a robust transcriptomic signature of TGFβ-induced EMT, and then used this signature to distinguish cancer cell lines and patient samples with evidence of TGFβ-induced EMT. We examine the signature across multiple breast cancer and pan-cancer datasets, demonstrating that: our results are reproducible on independent data; cell-line and patient samples show consistent, cancer type-specific levels of TGFβ-EMT activity, and; our TGFβ-induced EMT signature is influenced by the accumulation of genetic mutations across the TGFβ signalling pathway. Finally, we apply our signature to stratify patients and show differences in survival outcome, and identify cell lines with resistance to common cancer drugs.

Poster ID: 4 (AB3ACBS)

Simplifying simulation of single-cell RNA sequencing

Luke Zappia (a,b), Belinda Phipson (a) and Alicia Oshlack (a,b)

(a) Bioinformatics, Murdoch Childrens Research Institute; (b) School of Biosciences, The University of Melbourne

Single-cell RNA sequencing (scRNA-seq) is rapidly becoming a tool of choice for biologists who wish to investigate gene expression, particularly in areas such as development and differentiation. In contrast to traditional bulk RNA-seq experiments, which measure expression averaged across millions of cells, single-cell experiments can be used to observe how genes are expressed in individual cells. Along with the dramatic increase in resolution provided by scRNA-seq comes an array of bioinformatics challenges. Single-cell data is relatively sparse (for both biological and technical reasons), quality control is difficult and it is unclear how to replicate measurements. The focus of analysis is also different, with more emphasis on clustering cells to identify cell types or ordering of cells to understand dynamic processes than traditional tasks such as differential expression testing. Any new bioinformatics method for scRNA-seq analysis should demonstrate two things: 1) it can do what it claims and 2) it helps to produce biological insight. The first is hard to prove on real data where there is often no known truth. Because of this, bioinformaticians turn to simulations. Unfortunately current scRNA-seq simulations are frequently poorly documented, not reproducible and do not demonstrate similarity to real data or experimental designs. Here we discuss some of the problems with simulating scRNA-seq data and provide a simulation framework that addresses these concerns.

Poster ID: 5 (AB3ACBS)

Accurate and robust normalization of Nanostring nCounter gene expression data

Ramyar Molania(a,b), Terence P Speed(c,d), Alexander Dobrovic(e,f)

(a) Department of Medicine, Austin Health, University of Melbourne; (b) Translational Genomics and Epigenomics Laboratory, Olivia Newton-John Cancer Research Institute; (c) Bioinformatics Division, Walter and Eliza Hall Institute of Medical Research; (d) Department of Mathematics and Statistics, University of Melbourne; (e) Department of Pathology, University of Melbourne; (f) School of Cancer Medicine, La Trobe University

Nanostring nCounter is being increasingly used for research and clinical studies due to its capability to directly measure gene expression for a large panel of genes from RNA, even when the RNA is substantially degraded. Most suggested normalization methods including the Nanostring nSolver software are based on using small panels of user-chosen housekeeping genes and Nanostring spike-in controls. These are insufficient to remove batch and other technical effects, particularly when the data comes from large or complex experiments.

We used a new normalization method, Removing Unwanted Variation III (RUVIII), which is a variant on the RUV method using replicate samples as published by Jacob et al, Biostatistics 2015. We compared the performance of RUVIII with other normalization methods using several evaluation criteria including similarity of replicate samples, box and Relative Log Expression (RLE) plots of un-normalized and normalized expression values, differential expression analyses and clustering methods including tSNE and PCA.

We applied the RUVIII method to two different in-house data sets, as well as 8 different studies with nearly 8000 samples from publicly available repositories. The results show the performance of RUVIII is markedly better than existing methods. In particular, the performance of RUVIII in dealing with complex experiments with substantial batch effects is significantly superior compared to other methods. RUVIII not only removes batch effects in the data sets, but also reveals otherwise masked biology in the experiments.

In conclusion, removing unwanted variation plays an essential role in obtaining precise results for gene expression data. Using panels of housekeeping genes to normalize Nanostring gene expression data may lead to unsatisfactory normalization and accordingly, misleading results. RUVIII gives superior performance as judged both metrics and by concordance with known biological findings.

Poster ID: 6 (AB3ACBS)

Investigating the 3’ untranslated region of mRNA in order to understand the drivers of metastasis in primary triple negative breast tumours

Andrew Pattison (a), Cameron Johnstone (a,b), Paul Harrison (a,c), Robin Anderson (b), David Powell (c) and Traude Beilharz (a)

(a) Department of Biochemistry and Molecular Biology, Monash University, Melbourne, Australia; (b) Metastasis Research Laboratory, Olivia Newton-John Cancer Research Institute, Melbourne, Australia; (c) Monash Bioinformatics Platform, Monash University, Melbourne, Australia

While progress has been made in the detection of primary breast tumours, determining how likely a tumour will be to metastasise is still very difficult. Understanding this “metastatic potential” is most important in the so called “triple negative breast cancers” (TNBCs) which lack the classical markers that are commonly targeted in treatment. Chemotherapy is usually given for triple negative tumours, often unnecessarily. Better markers of tumour metastatic potential are clearly required.

Alternative polyadenylation (APA) is the process whereby the poly (A) tail is added to the 3’ untranslated region (3’ UTR) of mRNA at one of multiple possible sites, changing 3’ UTR length and potentially the regulatory elements that bind to it. APA has been shown to be indicative of tumour state, but is often overlooked when conducting RNA-seq analysis. We are developing a method to cheaply and effectively quantify the expression state of a primary breast tumour based off Poly (A) Test sequencing (PAT-seq) data, which sequences 3' UTRs in a genome wide fashion. PAT-Seq is also capable of measuring differential poly (A) tail length which may also play a role in metastasising tumour cells.

We hypothesise that the metastatic potential of a primary tumour can be calculated from changes in gene expression, APA site usage and the length of the poly (A) tail. We are testing this hypothesis in an increasingly metastatic cell line model both in vitro and in vivo. The model was generated from MDA-MB-231 TNBC cells and differences should be associated with metastatic potential. We have discovered some interesting 3' UTR associated changes in immune signalling, calcium ion balance and glutamine metabolism genes.

In order to effectively interpret and visualise PAT-seq data we make use of “Tail-tools”, a custom bioinformatics pipeline that employs modified versions of Limma and Degust. Recently, we have also begun to use Shiny to visualise PAT-seq data in new ways, often on a per gene basis.

Poster ID: 7 (AB3ACBS)

Insertion sequence elements are drivers of diversification in the broad host range aquatic pathogen Streptococcus iniae

Areej Alsheikh-Hussain (c), Nouri L. Ben Zakour (a), Andrew C. Barnes (b) and Scott A. Beatson (c)

(a) The Westmead Institute for Medical Research, Sydney, Australia
(b) School of Biological Sciences and Centre for Marine Science, University of Queensland, Brisbane, Australia
(c) School of Chemistry and Molecular Biosciences, Australian Centre for Ecogenomics, University of Queensland, Brisbane, Australia

Fish mortality caused by Streptococcus iniae is a major economic problem in fresh and seawater finfish aquaculture in warm and temperate regions globally, including Australia and Southeast Asia. There is also risk of occasional zoonotic infection by S. iniae through handling of contaminated fish. Vaccination of fish against S. iniae sometimes fails due to the rapid mutation of genes in the capsular biosynthesis operon of S. iniae, which leads to novel serotypes and consequent re-infection of immunized stock. Currently, genomic analysis of S. iniae is scant, with only 4 complete genomes sequenced to date, although high strain diversity has been described using pulsed-field gel electrophoresis of genomic DNA. The role of mobile genetic elements, including insertion sequences (IS), in this diversification of S. iniae lineages or as mechanisms of adaptation has not yet been investigated. The aim of the present study is to investigate the evolutionary history of S. iniae and its adaptation mechanisms that enable such widely disseminated infection in different fish species, environments, and humans. To achieve this we sequenced a collection of 113 S. iniae isolates from different hosts of worldwide origin. A phylogenetic tree constructed by maximum likelihood analysis of alignment of non-recombinant core-genome SNPs from 113 strains in the collection, along with the four complete genomes available on Genbank revealed separate clustering of human and fish isolates, as well as phylogeographic grouping even within Australia, with strains from Queensland clustering separately to those from the Northern Territory. The distribution of IS elements was found to be clade-specific, where the CRISPR/Cas system was targeted by different IS types causing interruption of cas9, cas1, and csn2, some of which are homoplasies. This broad scale assessment of diversity and diversification at the genomic level forms the basis for identification of stable, conserved potential antigens for future development of improved vaccination against S. iniae and this work is now underway.

Poster ID: 8 (AB3ACBS)

Elucidating the tolerance of ammonia oxidising bacteria to Free Nitrous acid (FNA) using a combined metagenomic and quantitative MS-SWATH metaproteomic approach.

A E Laloo(a), J Wei(a), D Wang(b), S Narayanasamy(c), A Buschart(c), Q Wang(a), I Vanwonterghem(d), Jason Steen(d), B Schultz(e), P Wilmes(c), A Nouwens(e), P Hughenholtz(d), Z Yuan(a), P L Bond(a).

(a) Advanced Water Management Centre (AWMC), University of Queensland, Australia
(b) College of Environmental Science and Engineering, Hunan University, Changsha 410082, China.
(c)Luxembourg Centre for Systems Biomedicine, Université du Luxembourg.
(d) Australian Centre for Ecogenomics, University of Queensland, Brisbane, Australia.
(e) School of Chemistry and Molecular Bioscience, University of Queensland.

Acidified Nitrite or Free Nitrous acid (FNA) is known to have strong inhibitory and biocidal effects in parts per billion (ppb) and parts per million (ppm) levels. Most micro-organisms have very low tolerance to FNA, however among nitrifiers, ammonia oxidising bacteria (AOB) rather than nitrite oxidising bacteria (NOB), this despite NOB having more pathways available for nitrite detoxification. Herein we investigate the reasons for this atypical behaviour in AOB using a quantitative metaproteomics method in an enriched activated sludge. On investigation we found that FNA tolerance in AOB is primarily due to the up regulation of the oxidative stress enzymes such as peroxidases and reductases that remove reactive nitrogen species in response to oxidative stress caused by FNA. Enzymes involved in ATP production such as ammonia monoxygenase (amo) and hydroxylamine oxidoreductase (hao) were upregulated suggesting that ATP could be used for energy dependent detoxification mechanisms such as the conversion of nitrite to ammonia as evidenced from the quantitative metaproteomics. This study provides the first detailed quantitative metaproteomic approach to study the antimicrobial effect of Free Nitrous acid on AOB and NOB found in activated sludge.

Poster ID: 10

De novo assembly of Sugarcane leaf RNAseq data: A cluster and merge approach

Hertweck, Kate (a), Wathen-Dunn, Kate (b, c)

(a) University of Texas at Tyler
(b) Sugar Research Australia
(c) University of Queensland

Yellow Canopy Syndrome (YCS) is currently an undiagnosed plant syndrome affecting Sugarcane crops throughout Queensland. The symptoms include reduced photosynthesis and transpiration, accumulation of sugars and starch in the leaves, and the appearance of yellow colouration in mid-canopy leaves, usually on one side of the leaf blade and beginning mid-leaf then spreading outwards.

As most of the symptoms occur in the leaf, we are using the differential expression in the leaf transcriptome to investigate possible causes for the syndrome, comparing YCS-symptomatic leaves with Control non-symptomatic leaves.

We have RNAseq data from 70 leaf samples, taken at different times, from different varieties and from different growing regions. As Sugarcane and related species currently lack a high-quality reference genome or EST database, we performed a de novo assembly of all 70 samples using multiple algorithms and multiple kmers.

Here, we describe the results from various cluster and merge approaches used to generate a single Reference Transcriptome from multiple kmer assemblies of four different algorithms: Trinity, Velvet, SOAP2 and SOAPTrans.

Poster ID: 12

Predicting Motif Mimicry in Viruses

Sobia Idrees (a), Åsa Pérez-Bercoff (a), Richard J Edwards (a)

(a) School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, Australia

Many viruses hijack host cellular machinery through mimicry of Short Linear Motifs (SLiMs) that interact with host protein domains. SLiMs are short stretches of amino acids (~3-10) which are involved in post translational modifications (PTMs), protein-protein Interactions (PPIs), cell regulation and cell compartment targeting. To date, several studies have been conducted to identify PPIs, but no specific study to see how well different PPI-capturing methods capture SLiMs-mediated interactions. The main objectives of this study are:1) to predict Domain Motif Interactions (DMIs)among viral and host proteins;2) to find whether virus-human PPIdata from the virhostome resource is enriched for DMIs; and 3) to see which PPI method is better for studying DMIs. Results have shown that virhostome data is enriched for DMIs and can be a good source to study motif mimicry in viruses. Permutation tests showed more enrichment for DMI in TAP data than Y2H data. Moreover, novel candidate DMIs have been discovered which need further validations. The outcome of this study will be helpful in uncovering unique strategies of viruses to interact with human proteins which will eventually be significant for pathogen research.

Poster ID: 13

Dimensionality reduction by t-SNE uncovers the connections that shape the landscape of the transcriptome

Michael See (a), Paul F Harrison (b) David R Powell (b) David Albrecht (a) and Traude Beilharz (c)

(a) Faculty of Information Technology, Monash University, Melbourne, VIC 3800, Australia
(b) Monash Bioinformatics Platform, Monash University, Melbourne, VIC 3800, Australia
(c) Development and Stem Cells Program, Monash Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology, Monash University, Melbourne, VIC 3800, Australia.

The increasing use of high throughput sequencing technologies has resulted in the ever growing size of international biological data repositories such as Gene Ontology Omnibus (GEO), Saccharomyces Genome Database (SGD) and The Cancer Genome Atlas (TCGA). These databases contain a large amount of underutilised high dimensional data from thousands of biological experiments. We are specifically interested in revealing the underlying biological connectivity in gene-expression networks within this resource.

To make use of these data, we employ the Exploratory Data Analysis (EDA) paradigm which is characterised by a use of visualisations to suggest possible hypotheses which would otherwise remain undiscovered by standard statistical hypothesis testing. As the high-content biological data are difficult to interpret in its native state, we use a dimensionality reduction algorithm to produce visualisations of multi-assay data of both experimental and existing gene expression data.

We show that using t-Stochastic Neighbour Embedding (t-SNE) algorithm as an approach to feature extraction, it is possible to identify functional aspects of the transcriptomic landscape which we can then verify using Gene Ontology associations.

Poster ID: 14

Quantification of paediatric burn blister fluid proteome using SWATH MS to assist better clinical diagnosis

Tuo Zang (a,b,c), Daniel A. Broszczak (a,b,c), James A. Broadbent (a,b,c), Leila Cuttle (a,b,d), Tony J. Parker (a,b)

(a) Tissue Repair and Regeneration Program, Institute of Health and Biomedical Innovation, Queensland University of Technology, Kelvin Grove, Australia.
(b) School of Biomedical Sciences, Faculty of Health, Queensland University of Technology, Brisbane, Australia.
(c) Wound Management Innovation Co-operative Research Centre, Brisbane, Australia.
(d) Centre for Children's Burns and Trauma Research, Queensland University of Technology, Institute of Health and Biomedical Innovation at the Centre for Children's Health Research, South Brisbane, Australia.

Burn injury is a common and traumatic event in paediatric population. At present, the diagnosis of burn injury severity is largely dependent on the clinician's experience which is time consuming and delays the timely treatment. Objective and quantitative measures to aid in diagnosis are absent from clinical practice due to a lack of information on the biochemistry of burn injury. Burn blister fluid (BF) is considered to be a viable source of biomolecules that reflect relevant systemic responses and the local microenvironment. A more comprehensive understanding of the biochemistry of burn wounds will open new avenues for diagnostic and prognostic development.

In order to generate a comprehensive protein inventory, some of the BF samples were pooled according to burn depth (12 superficial (S), 12 deep-partial thickness (D), and 4 full thickness (F)) and fractionated by four different methods, including filter based, gel electrophoresis, isoelectric focusing and immuno-depletion. The samples were subsequently processed by standard digestion protocols. Peptides were then analysed using liquid chromatography tandem mass spectrometry (LC-MS/MS). Following this, individual BF samples (n=100) were processed by in-solution digestion and the tryptic peptides were analysed using LC-MS/MS in sequential window acquisition of all theoretical mass spectra (SWATH) mode to obtain quantitative data.

More than 800 individual proteins were identified and formed the basis of a BF protein library. Relative abundances of more than 600 proteins in every individual sample were extracted. Subsequent gene ontology enrichment analysis revealed underlying biological processes respond to different burn severity. Some statistical and bioinformatic analysis were utilized to discover the proteins can classify burn depths.

In this study, the biological processes at initial stage of burn injury and its relation to different burn depth were quantitatively profiled. Therefore, it contributes to the knowledge of burn wound microenvironment biochemistry. Some specific proteins have been found that can be used to classify burn severities.

Poster ID: 15

Differences in non-seed pairing between miR-324-3p and miR-1913-3p account for binding and functional differences

Sarah M Williams (a,b), Belinda J Goldie (c), Janette Edson (d), Michelle Watts (b,d), Charles Claudianos (b,d), Alexandre Cristino (a)

(a) Diamantina Institute, University of Queensland
(b) Monash University
(c) University of Newcastle
(d) Queensland Brain Institute, University of Queensland

The binding of microRNAs to target sites on the 3' untranslated regions (UTRs) of target genes is affected by many factors. The presence of an exact 'seed' complementary match from bases 2-7 is often very important, but complementarity in other regions of the mature microRNA sequence also plays a role. MicroRNAs miR-324-3p and miR-1913-3p are mature microRNAs that share a seed sequence, but have different mature sequences overall. While miR-1913 has evolved recently within primates, miR-324 is more broadly conserved. We were interested in the degree of similarity between the binding patterns of these microRNAs, and if differences in targeting could indicate functional divergence. We used biotin-tagged microRNA mRNA-pulldown assays to measure the binding profiles of those microRNAs, and also measured the effect of exogenous overexpression (via transfection) of these microRNAs in the cells. The results indicate a general association of miR-324-3p and miR-1913-3p with cellular differentiation, consistent with their expression changes seen during differentiation of SH-SY5Y human neuronal derived cell line, lending further support to an anti-proliferative effect described for miR-324 in the literature. To examine the difference between miR-324-3p and miR-1913-3p targeting, we examined their relative bias to communities in a mRNA coexpression network and protein-protein interactions networks. This revealed a miR-1913-3p-specific association with translational processes and the proteasome, in contrast to the more proliferation and cancer related pathways most prominent for miR-324-3p. More generally, these results provide some characterisation of the targeting and function for these two microRNAs, particularly miR-1913 for which there is little information available, and indicate a possible importance in neuronal function.

Poster ID: 17

Assembly challenges from bacterial mobile genetic elements

Anna Syme, Sarah Baines, Dieter Bulach, Torsten Seemann

VLSCI and MDU, The University of Melbourne

Bacterial mobile elements are plasmids, transposons, IS elements, integrons and phage. They can jump in and out of plasmids and chromosomes, move between bacteria, and often carry genes that confer resistance to antibiotics. Antibiotic resistance can therefore spread quickly through bacterial populations, even to bacteria from different species. We need to identify these mobile elements by correct assembly from sequencing reads. Being small, repetitive, and mobile causes assembly challenges. Here we present a workflow for this task, being developed as part of a national Australian project to sequence and examine the genomes of bacteria involved in blood poisoning. The workflow includes hybrid pacbio-illlumina assembly and uses the tools Canu, Circlator, Spades, Bandage and Pilon.

Poster ID: 18

In silico analysis of immunomodulatory vaccine candidate proteins SpyCEP and EndoS in Streptococcus pyogenes

Lochlan Fennell (a,b)

(a) University of the Sunshine Coast
(b) QIMR Berghofer Medical Research Institute

BACKGROUND: Group A Streptococcus is a pathogenic bacterial species responsible for a plethora of human infections. Vaccine development for the pathogen has focused primarily on the M protein. Given the hyper-variability of the N terminus of the protein, multi-antigen vaccines have emerged as a potential vaccine strategy. SpyCEP and EndoS are immunomodulatory proteins and form two potential vaccine candidates. In this study we assessed their suitability as vaccine candidates using a number of in-silicon techniques. METHODS: 51 complete S. pyogenes genomes were obtained from the NCBI genome database. SpyCEP and EndoS sequences were isolated using the BLAST+ suite. Sequences were aligned using MUSCLE. Phylogenetic analysis was performed using Mega7. Protein structure and modelling was conducted using SWISS-MODEL. Hydrophobicity and antigenicity were predicted using the Kyte-Doolitte and Kolaskar-Tongankar window methods. RESULTS: EndoS and SpyCEP had high overall sequence identity. Phylogenetic analysis revealed that protein sequence identity was highly correlant with serotype. Combinatorial analysis of DN/DS ratios, hydrophobicity and antigenicity predictions indicated residues 960-980 of SpyCEP as the most robust and suitable residues for vaccine inclusion, while residues 420-465 represent the highest predicted vaccine suitability score. CONCLUSION: Our in-silico analysis supported the use of SpyCEP and EndoS peptides in multiple antigen vaccine strategies. Moreover we were able to predict potential robust and suitible regions of both SpyCEP and EndoS for inclusion in peptide based polyvalent vaccines.

Poster ID: 19

Dissecting the taurine-indicine balance in fertility-related bovine protein coding regions

Parthan Kasarapu (a), Laercio R. Porto-Neto (a), Marina R. S. Fortes (b), Sigrid A. Lehnert (a), Mauricio Mudadu (c), Luiz Coutinho (d), Luciana Regitano (c), Andrew George (e), Antonio Reverter (a)

(a) CSIRO Agriculture and Food, Queensland Bioscience Precinct, St. Lucia, QLD 4067, Australia
(b) School of Chemistry and Molecular Biosciences, The University of Queensland, Australia
(c) Embrapa Southeast Livestock, Rodovia Washington Luiz, Km 234, Sao Carlos, SP, Brazil
(d) University of Sao Paulo, Brazil
(e) DATA61, Brisbane QLD 4001, Australia

The analysis of genomic data provides insights into the biodiversity in organisms. Specifically, in beef cattle, genomic information represented using single nucleotide polymorphism (SNP) data can be mined to infer characteristics of cattle breeds. We present a bioinformatics pipeline for the exploratory analysis on a large SNP database of cattle driven by heterozygosity and Hardy-Weinberg equilibrium (HWE) values and investigate new means to identify molecular pathways under selection. We perform analysis on a comprehensive data set of 18,363 cattle from 19 breeds with genotypes for 729,068 SNPs. We selected the SNPs that are mapped to autosomal chromosomes and within 1 kb of an annotated protein coding gene. The edited data contained 246,864 SNPs mapped to 8,631 genes. We clustered the SNPs at the gene level based on their heterozygosity and identified SNPs that deviate from the HWE. The resulting clusters separate breeds by lineages with pure Bos indicus and Bos taurus breeds at the extremes, while cross breeds and tropically adapted composites appear as distinct clusters. In addition, Gene Ontology enrichment analysis of genes ranked on their heterozygosity demonstrate the discriminative power of heterozygosity in detecting pathways of interest. Furthermore, we mined publicly available databases to classify our genes into four functional categories of transcription factors (TF), tissue-specific, secreted proteins, and kinases. With a filtered set of 1,259 genes we developed a co-heterozygosity network that identified 47 TF playing a central role in the taurine-indicine axis. A further focus on genes that impact fertility phenotype in beef cattle brings new insights about the association of heterozygosity with fertility in tropically-adapted cattle breeds. Our pipeline can be employed to any scenario where population structure deserves scrutiny at the molecular level particularly in the presence of a prior set of genes known to impact a particular phenotype.

Poster ID: 20

Finding optimal coverage

Anna Quaglieri (a,b), Terry Speed (a), Ian Majewski (a)

(a) Walter and Eliza Hall Institute
(b) University of Melbourne

Next-Generation Sequencing (NGS) technologies offer the possibility of sequencing large amount of genetic material from several samples at the same time, revolutionising the way we study diseases. However, the experimental design is often overlooked resulting in suboptimal statistical power and high financial costs.

In a NGS experiment the coverage, seen as the average number of times that a base of a genome is sequenced, and the number of samples are fundamental factors affecting both the costs and the results of an experiment. The choice of coverage is especially critical in cancer genomics where normal tissue is mixed with tumour and some cancer mutations appear with a very low frequency. At the same time power calculations are not trivial.

We have the opportunity to plan the sequencing design of a set of genomic samples from a cohort of Core Binding Factor Acute Myeloid Leukemia (CBF-AML) patients collected by the Australasian Leukemia and Lymphoma Group as part of a controlled trial. The clinical focus of the research is to characterise the mechanism behind relapse in adult CBF-AML which still affects roughly the 40% of the patients treated by standard chemotherapy.

To inform the planning of the sequencing design within the available budget, we implemented a preliminary analysis using forty-six publicly available CBF-AML RNA-seq samples produced by the Leucegene group. The Leucegene data were sequenced at a very high coverage and the results of their analysis have just been published (Lavallée et al. 2016).

Using their results as a benchmark, we plan to study how much information is lost by randomly and sequentially reducing the coverage of the original data. This is going to help us to design a cost-efficient and powerful study as well as inform other genomic study designs.

Poster ID: 21

In silico functional characterization of the human lipid raft proteome

Anup D. Shah (a), David Chen (b), Melissa J. Davis (c) and Michelle M. Hill (a)

(a) The University of Queensland Diamantina Institute, The University of Queensland, Translational Research Institute, Brisbane, Queensland, 4102, Australia
(b) School of Information and Communication Technology, Griffith University, Brisbane, QLD, Australia 4111
(c) Division of Bioinformatics, The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville Victoria 3052, Australia

Cholesterol and sphingolipids in the cellular membranes form functional microdomains that serve as a dynamic protein sorting and signaling platforms. These specialized membrane regions, also called lipid rafts, regulate numerous cellular events and their dysfunction is implicated in multiple diseases. Despite their functional importance, the molecular features that facilitate recruitment of proteins to lipid-rafts are poorly understood. According to the findings from low-throughput/targeted biophysical investigations, proteins can be sorted to the lipid raft via a number of mechanisms such as lipid attachment or binding, membrane penetration or protein-protein interactions. To determine the relevance of these features, this chapter investigates the representation of these structural determinants in RaftProt 2.0, a comprehensive human raft proteome dataset of 4898 proteins, derived from 75 proteomics experiments from 54 different cell and tissue types. Using a detailed computational analysis, the human raft proteome was classified into intrinsic proteins, extrinsic proteins and co-isolated proteins, based on the potential molecular mechanism of raft partitioning. Of all the suggested raft-targeting mechanisms, our meta-analysis suggested that the palmitoylated and the cholesterol binding proteins are selectively enriched in the human raft proteome compared to the neighboring plasma membrane on the cell surface. Protein-protein interaction network analysis of raft proteins suggest a high degree of interconnections among lipid raft proteins. Furthermore, functional and pathway enrichment analysis revealed key biological processes regulated by lipid rafts. Collectively, the global integrative analysis presented here will pave the way to more refined representation of human lipid raft associated proteins, which will allow better understanding of their function and regulation.

Poster ID: 22

Machine-learning annotation of Human splicing branchpoints

Beth Signal (a,b), Brian S Gloss(a,b), Marcel E Dinger(a,b), Tim R Mercer(a,b)

(a) Genomics and Epigenetics, Garvan Institute of Medical Research, Sydney, Australia
(b) St Vincent's Clinical School, University of New South Wales, Sydney, Australia

The branchpoint is a basal genetic element required for gene splicing. Despite a primary role in exon inclusion, current annotations of branchpoints are incomplete and limited to experimental catalogues. Due to difficulty in experimentally identifying branchpoints compared to other splicing elements, their contribution to normal and alternative splicing has been largely understudied. We have developed a machine-learning algorithm, using the most recent gold-standard human branchpoint annotations, to identify branchpoints from gene sequence alone. Using this approach, we are able to locate branchpoints in 85% of introns in current gene annotations. This near-complete annotation is unbiased towards gene type and expression levels, highlighting the advantage of developing predictive models in genome annotation. Several introns were found to encode multiple branchpoints, which may be a mechanism through which mutational redundancy is encoded in key genes. Branchpoint strength is associated with differing modes of alternative splicing, and has a distinct contribution to other splice elements. Notably, this annotation constitutes an invaluable resource for interpreting the mechanistic impact of common- and disease-causing human genetic variation on gene splicing.

Poster ID: 23

Leveraging uncertainty in ancestral sequence reconstruction using partial order graphs.

Gabriel Foley

School of Chemistry and Molecular Biosciences, University of Queensland, Brisbane, Queensland 4072, Australia

Reconstructing the ancestors of sets of aligned proteins provides information about conserved positions and shared evolutionary histories. This information can be used to identify positions within ancestral proteins with the potential to be mutated to improve protein function while preserving structural integrity. Once an ancestor and amenable sites have been identified, synthetic DNA templates can be designed that allow for generation and characterisation of ancestral variants.

The problem of uncertainty manifests itself in two distinct areas and with a multitude of effects. The first area is the initial protein alignment, as increasing the number and diversity of sequences increases the presence of insertion and deletion events and negatively impacts our ability to accurately align proteins. The second area is the generation of ancestral variants, as predictions of individual positions that have close to equally likely probabilities give rise to a huge number of potential ancestors that are impossible to comprehensively characterise.

Our work looks to use partial order graphs to control both of these areas of uncertainty. Firstly, by using graphs for protein alignment to represent ambiguity and to consider a greater amount of information when classifying an alignment event as either an insertion or deletion. Secondly, by using the fully populated alignment graphs as probabilistic models we extensively sample from and enabling us to rank the likelihood of ancestral variants. This allows the physical generation of proteins to focus on those identified as more likely to fold and function.

We are currently extending and improving partial order graphs by adopting techniques successfully employed in other sequence alignment programs such as the Maximum Expected Accuracy algorithm and the use of ancestors as a measure of probabilistic consistency. We are also developing methods to sample from the aligned graph and preference ancestral variants with combinations of amino acids more commonly observed in nature.

Poster ID: 24

Cancer progression and hypoxia - development of a pan cancer hypoxic-signature

Kristy Horan, Sepideh Foroutan, Melissa Davis

Bioinformatics Division Walter and Eliza Hall Institute Royal Parade Melbourne Victoria

During tumorigenisis, tumor mass reaches a critical point at which areas of the tumor are deprived of oxygen, for a short period (acute) or extended period (chronic). This leads to a localised condition termed hypoxia. Growing evidence suggests that, although hypoxia can lead to cell death, it can also promote cancer progression and metastasis via increasing vascularisation at the tumor site, increasing metastasis to secondary sites via enhancement of epithelial-mesenchymal transition (EMT), altering the cells dependence on nutrients and altering the metabolism of drugs. Recently, it has been demonstrated that the response of breast cancer cells to anti-cancer drugs was markedly different in cells under hypoxic conditions, with the drugs in fact increasing proliferation.

The aim of the current study is to identify a pan cancer hypoxic signature. We have identified nine appropriate studies with a range of cancer cell lines from publicly available micro array data sets from Gene Expression Omnibus and Array Express. We are using a meta analysis approach (product of rank), to identify relevant genes under acute and hypoxic conditions. This signature may then be applied to patient data sets to investigate correlations between primary tumor size, metastasis and, where data is available, drug response or resistance. In addition, it may provide a prognostic tool to allow more appropriate targeting of drugs and treatments in patients.

Poster ID: 25

Outbreak of carbapenem-resistant Acinetobacter baumanii (CRAB) in a Brisbane Intensive Care Unit

Leah W. Roberts (a,b), Patrick Harris (c,d), Brian M. Forde (a,b), Graeme Nimmo (d), Narelle George (d), Krispin Hajkowicz (e), Jeff Lipman (f), Mark A. Schembri (a,b), David Paterson (c, e), Scott A. Beatson (a,b)

(a) School of Chemistry and Molecular Biosciences
(b) Australian Infectious Disease Research Centre, University of Queensland, Brisbane QLD
(c) UQ Centre for Clinical Research, Brisbane QLD
(d) Pathology Queensland, Central Laboratory, Brisbane, QLD
(e) Unit of Infectious Diseases, Royal Brisbane and Women's Hospital, Herston, Queensland, Australia
(f) Burns Trauma and Critical Care Research Centre, The University of Queensland Australia.

Introduction:

Acinetobacter baumannii is an important nosocomial pathogen of critically ill patients that has become increasingly hard to treat due to rising antibiotic resistance. Here we undertook a genome-led investigation of a carbapenem-resistant A. baumannii (CRAB) outbreak in an intensive care unit (ICU) in Brisbane.

Methods:

28 CRAB isolates from 17 patients were collected between May-August 2016 and sequenced using the Illumina platform. Many of the patients displayed co-infection with Klebsiella pneumoniae, Serratia marcescens and Pseudomonas aeruginosa, which were also sequenced. Rapid antibiotic resistance gene profiling and multi-locus sequence typing (MLST) were performed using the Nullarbor pipeline. Core genome single nucleotide polymorphisms (SNPs) were determined using Nesoni.

Results:

All CRAB isolates were found to be sequence type (ST)1050 and differed by less than 13 core SNPs, indicative of direct transmission within the hospital. Sequencing of earlier CRAB isolates in the same hospital from March-May 2016, 2015 and 2000-2006 found no close relationship with the current CRAB outbreak, suggesting that the index case had been introduced into the hospital. Carbapenem resistance in the CRAB isolates was driven by both overexpression of the beta-lactamase OXA-23 by ISAba1 and overexpression of chromosomal ampC by ISAba125. The CRAB isolates were also found to carry the transposon Tn6279, conferring resistance to macrolides, aminoglycosides and chloramphenicol. Horizontal transferal of antibiotic resistance genes was seen between K. pneumoniae and S. marcescens isolates, highlighting the easy spread of resistance genes within a short time span.

Conclusions:

Genome sequencing was able fully characterise a CRAB outbreak in a Brisbane ICU and determine transmission pathways between critically ill patients and mechanisms of antibiotic resistance. Overall these results highlight the potential of genome sequencing in the clinical microbiology setting.

Poster ID: 26

Biomarkers of Anterior Cruciate Ligament Injury and Recovery

Yee L Chng (a,b), Anthony W Parker (c), David A Parker (d), Tony J Parker (a,b)

(a)Tissue Repair and Regeneration Program, Institute of Health and Biomedical Innovation, Queensland University of Technology, Brisbane, Australia
(b)School of Biomedical Sciences, Faculty of Health, Queensland University of Technology, Brisbane, Australia
(c)Injury Prevention Program, Institute of Health and Biomedical Innovation, Queensland University of Technology, Brisbane, Australia
(d)Sydney Orthopaedic Research Institute, Sydney, Australia

Anterior cruciate ligament (ACL) and associated knee injuries severely impact on an individual's opportunity to resume previous physical activity levels and may also increase the risk of developing knee osteoarthritis. While magnetic resonance imaging provides excellent diagnostic information on the specific structures involved in the injury, it is expensive and therefore not practical for providing repeated follow up information after ACL reconstructive surgery. Recent advances in protein profiling approaches including liquid chromatography – tandem mass spectrometry (LC-MS/MS) provides evidence to suggest that novel diagnostic biomarkers of knee injury may be detectable. Hence, the focus of this research project was to interrogate the urinary proteome, using LC-MS/MS and associated proteomics techniques, to reliably quantify and monitor the structural damage following acute ACL injury and repair status after surgery. Pre- and post-operative urine samples from 14 confirmed ACL injured patients and 11 healthy control subjects were analysed by LC-MS/MS using a TripleTOFTM 5600+ with SWATHTM acquisition, coupled to an Eksigent nanoLC system. The MS data were analysed using ProteinPilot, PeakView and MarkerView (AB Sciex) and resulted in the generation of a spectral library for more than 500 proteins. Multivariate statistical analysis revealed several proteins and fragments that were associated with the injured cohort and the proteins that could be indicative of tissue recovery outcomes. In addition, significantly overrepresented biological process, cellular component and molecular function gene ontologies related to the injured compared to the non-injured group were also determined. The identification and quantification of proteins relating to tissue damage may provide the basis for monitoring the tissue recovery following surgery. Such biomarkers may provide additional tools for clinicians in the evaluation of both the initial injury but importantly, tissue recovery following surgery.

Poster ID: 27

eQTL analysis of quantitative endophenotypes for ocular health in the Norfolk Island isolate

Pik Fang Kho, Rodney Lea, Miles Benton, David Eccles, Larisa Haupt, Alex Hewitt, David Mackay and Lyn Griffiths

Genomics Research Centre, Institute of Health and Biomedical Innovation, Queensland University of Technology, Brisbane, Queensland, Australia

Ocular health is influenced by both genetic and environmental factors. Despite significant progress in identifying genetic variants associated with ocular disorders there remains a substantial amount of unexplained heritability. Study design features that may assist in characterising the unknown genetic architecture includes; focusing on multiple quantitative traits related to ocular disorders (ie. endophenotypes), targeting genetic variants that directly influence gene expression (ie. cis-eQTLs) and utilising genetically isolated populations to reduce genetic and environmental noise and thus enhance association signals. In this study we performed heritability analysis of 14 ocular health endophenotypes measured in ~600 individuals from the NI isolate. For all heritable traits we performed pedigree-based association analysis of 200 SNPs previously shown to influence gene expression in whole blood collected from this cohort (ie. eQTL SNPs). Finally we performed transcript-by-trait regression analysis conditioned on eQTL SNP to identify transcript|SNP association with ocular endophenotypes. Statistical significance thresholds were set to balance type I and type II error rates. Results of this study revealed 9 heritable ocular endophenotypes, with estimates ranging from 0.35 for intraocular pressure to 0.82 for central corneal thickness (P<0.05). The most significant eQTL SNP association was between optic disc size and LPCAT2 (P=0.0004). The only transcript|SNP model that was statistically significant was for BTN3A2 and optic disc size, which included rs853676 (beta=0.23,P=0.008) and transcript (beta = 0.23,P=0.03) for an overall predictive model of R2 = 0.37 (P=1*10-7) BTN3A2 encodes butyrophilin, which is part of the immunoglobulin superfamily. This finding suggests a role for an inherited immune component in ocular health with regard to optic disc size variation. This study also demonstrates an alternate approach to understanding the genetic basis of ocular disorders.

Poster ID: 28

Effect of Serum Concentration on the Proteome of Rat Bone Marrow-Derived Mesenchymal Stem Cells

Morgan Carlton (a, b, c); Yinghong Zhou (c); Daniel Broszczak (a, b); Yin Xiao (c); Tony Parker (a, b)

(a) Tissue Repair & Regeneration Program, Institute of Health and Biomedical Innovation, Queensland University of Technology, Brisbane, Queensland, Australia
(b) School of Biomedical Science, Faculty of Health, Queensland University of Technology, Brisbane, Queensland, Australia
(c) Bone Group, Orthopaedics, Trauma and Emergency Care Program, Institute of Health and Biomedical Innovation, Queensland University of Technology, Brisbane, Queensland, Australia

Mesenchymal Stem Cells (MSCs) could be a potential treatment for multiple diseases and injuries due to their ability to replenish most cells in the human body. Translation of this requires cells to be maintained in their non-differentiated state, which remains a major challenge in the field. Optimal conditions for culturing these cells are now required; thus, in this study, rat MSCs cultured with three different supplemental serum concentrations were investigated using a proteomics approach. Bone marrow-derived rat MSCs were established in medium containing 10% serum. Cells were then starved prior to the application of media with different concentrations of serum (0%, 2% or 10%) and cultured for 24 hours. Cellular protein was collected and prepared for qualitative and quantitative (SWATH) mass spectrometry using standard techniques. Multivariate analysis (PCA, PLS-DA and oPLS-DA) revealed biochemical differences across treatments. Gene ontology (GO) enrichment analysis was used to determine the biological processes and molecular functions of cells in response to treatments. A protein library containing 803 proteins was generated, 58.03% of which were observed in all three treatments, while 8.47%, 8.47% and 1.49% were unique to 0%, 2% and 10% serum, respectively. Fewer proteins were detected in the 10% serum group compared to lower concentrations possibly due to high abundant protein induced ion suppression. Multivariate comparison of the abundance of all 803 proteins showed that 0% and 2% serum culture conditions were more similar compared to the 10% serum culture conditions. This similarity was also reflected in the GO enrichment analysis. Significantly over-represented GO terms associated with apoptosis were observed in the 0% and 2% groups but not the 10% group. Based on the biochemical data, there was a more pronounced similarity between the serum free and 2% treatments compared to the 10% treatment. However, it is clear that unique biochemical features exists within all three treatments and these features may provide the insight needed to monitor and maintain MSCs in long term culture.

Poster ID: 31

InsituNet, a Cytoscape app for network visualisation of in situ sequencing data

John Salamon (a), Xiaoyan Qian (c), Mats Nilsson (c), David Lynn (a,b)

(a) EMBL Australia Biomedical Informatics Group, Infection and Immunity Theme, South Australian Health and Medical Research Institute, Adelaide, Australia
(b) School of Medicine, Flinders University, Bedford Park, Australia
(c) Science for Life Laboratory, Department of Biochemistry and Biophysics, Stockholm University, Stockholm, Sweden

Gene expression studies typically homogenise samples before sequencing, discarding spatial information on where transcripts are expressed. In situ sequencing is a novel method to generate spatially-resolved, in situ RNA localization and expression data. Gene-specific barcodes allow data for up to 40 different transcripts/genes at an almost single-cell resolution to be generated in situ. The resulting images can therefore display the location and intensity of a million or more individual transcripts in a tissue section. Few methods currently exist to analyze and visualize the complex relationships that exist between these transcripts or identify how these transcriptional profiles change in different regions of the tissue or across different tissue sections. Here, we present InsituNet, an innovative new application that converts in situ sequencing data into interactive network-based visualisations, where each transcript is a node in the network and edges represent the spatial co-localization relationships between transcripts. InsituNet identifies co-localisations that occur between transcripts both significantly more, and less, than statistically expected, given the frequency of the transcripts in the tissue. An automated sliding window function allows the generation of networks representing each individual section of the tissue and these networks enable users to quickly and easily identify regions where the transcriptional profiles are altered (e.g. regions associated with pathology). Alternatively, the user can also select (irregularly-shaped) regions of interest in the section for comparison to other regions. One can also compare how the transcriptional network changes across different tissue sections (e.g. healthy vs. disease). Where multiple networks are constructed their layout is spatially synchronised to facilitate comparison. InsituNet has been developed for the popular Cytoscape platform and will be publicly available following publication.

Poster ID: 32

Combining high throughput sequencing data to improve identification of transcription factor binding

Alex Essebier (a) and Mikael Boden (a)

University of Queensland, Brisbane, Australia

Math1 is a transcription factor (TF) which plays a role in neural development and Medulloblastoma. TFs are known to bind to regulatory regions defined by accessible chromatin detected by DNaseI hypersensitivity (DHS), and epigenetic features such as histone-3 lysine-4 tri-methylation (H3K4me3) and histone-3 lysine-4 mono-methylation (H3K4me1) indicating enhancers and promoters respectively. The genes targeted by a TF can be detected using RNA-sequencing and perturbing the system by knocking-out the TF of interest, in this case: Math1. By combining these datasets with chromatin immunoprecipitation (ChIP-seq) data for Math1; we can select high confidence binding sites, identify epigenetic patterns distinguishing sites and improve detection of Math1 gene targets.

We gathered the above datasets in Cerebellar granule neurons (CGNs) at P5/7 to create a high confidence set of TF binding sites for Math1. To qualify, we required that Math1 ChIP-seq peaks overlap a regulatory region defined by DHS, H3K4me3 and H3K4me1. We identified 4,149 high confidence peaks from 8,804 total peaks and determined that 3,362 were bound to an enhancer >2,000bps from the nearest TSS indicating Math1's preference for binding enhancers.

Using our high confidence binding sites we identified 2,283 putative target genes of Math1. By requiring that a target gene show differential expression based on RNA-seq of a Math1 knock-out, we reduced this to 743 relevant target genes. These targets are linked to regulatory regions that are distal or proximal with distinct epigenetic features. Finally, using StringDB, DAVID and KEGG, we narrowed the genes of interest to <50, 3 of which were selected for experimental validation.

Through combining datasets, we identify a small set of significant target genes for Math1. We can extract relevant information about a TF and gain insight into why and how it binds, and what influence different binding 'modes' have on target gene expression.

Poster ID: 33

Genomic characterisation of E. coli ST101; an extraintestinal pathogenic clonal lineage

Melinda M. Ashcroft (a,b), Brian Forde (a,b), Minh-Duy Phan (a), Kate Peters (a), Kok-Gan Chan (c), Teik Min Chong (c), Wai-Fong Yin (c), Mark A. Schembri (a), Scott A. Beatson (a,b)

(a) Australian Centre for Ecogenomics, The University of Queensland, St. Lucia QLD, Australia
(b) Australian Infectious Disease Centre and School of Chemistry and Molecular Biosciences, The University of Queensland, St. Lucia QLD, Australia
(c) Division of Genetics and Molecular Biology, Institute of Biological Sciences, Faculty of Science, University of Malaya, 50603 Kuala Lumpur, Malaysia

Extra-intestinal pathogenic Escherichia coli (ExPEC) are one of the most common causes of urinary tract and bloodstream infections. ExPEC strains from sequence type (ST) 69, 73, 95 and 131 are well-characterised, globally disseminated clones, with ST101 a recently emerging multidrug resistant (MDR) clone. ST101 are frequently associated with carriage of the New-Delhi metallo-beta-lactamase (blaNDM) gene that confers resistance to last-line carbapenem antibiotics. However, despite the rapid emergence of E. coli ST101, very little is known about the genome or methylome of these important human pathogens. Here we present the first comprehensive genomic analysis of the E. coli ST101 lineage. Using Pacific Biosciences Single Molecule Real-Time Sequencing technology, we defined the complete genome of two E. coli ST101 isolates: E. coli MS6192 and E. coli MS6193. We show the most prevalent sublineage of ST101 strains are characterised by carbapenem resistance and a distinct mobile genetic element (MGE) and plasmid profile containing numerous antimicrobial resistance determinants and methyltransferase (MTase) genes. The phylogeny of E. coli ST101 revealed a single lineage, distinct from other E. coli strains within the B1 phylogroup. Additionally, we identified 12 DNA MTases, 9 of which are carried on the chromosome and 3 carried on each of the larger plasmids. Three MTases are conserved among all ST101 strains, with only 1 other shared across the majority of the lineage. Two novel type I MTases were identified with DNA recognition sites unique to the E. coli ST101 lineage. Furthermore, we identified a type I MTase conserved in E. coli and Klebsiella pneumoniae strains encoding blaNDM. These results indicate substantial variance in the genomic content of E. coli ST101 and the association of MGEs in defining clades within ExPEC lineages. The availability of these high quality, complete ST101 genomes will provide an important reference for understanding this MDR pathogen.

Poster ID: 35

Colorectal Cancer Atlas and FunRich: Discovery tools for integrated ‘omics’ data analysis

David Chisanga(a),Mohashin Pathan (b), Shivakumar Keerthikumar(b), Naveen Chilamkurti(a),Suresh Mathivanan (b)

(a) Department of Computer Science and Information Technology, La Trobe University, Bundoora, Victoria 3086, Australia
(b) Department of Biochemistry and Genetics, La Trobe Institute for Molecular Science, La Trobe University, Melbourne, Victoria 3086, Australia

In recent years, advancements in highthrouput data collection techniques such as mass spectrometry (MS) and next generation sequencing (NGS) have enabled the study of entire proteomes and genomes. This in turn led to a proliferation in both qualitative and quantitative data which now pose analytical challenges for biologists on how to gain clinically relevant insights into pathophysiology of disorders including cancer. To mitigate some of these challenges, we have developed Colorectal Cancer Atlas (CRC Atlas) (www.colonatlas.org), an integrated web-resource that catalogues proteomic and genomic data for colorectal cancer cell lines and tissues. Currently the database has a catalogue of 62,251 protein identifications, >8.3 million MS/MS spectra, >822,710 gene sequence variants, 1,631 Post Translational Modifications (PTMs) affected by sequence variants (out of 88,819 PTMs) and 351 pathways with sequence variants all of which have been derived from 209 and 13,711 colorectal cancer cell lines and tissues respectively. The database allows for the analysis of these data in the context of signalling pathways, protein-protein interactions, Gene Ontology terms, protein domains, post-translational modifications and the drug sensitivity of cell line.

Downstream functional enrichment analysis of the data catalogued in CRC Atlas can further be performed by FunRich, a desktop-standalone functional enrichment tool that we have also developed. Using FunRich, users can perform functional enrichment analysis for >13,000 species using UniProt and customizable database with gene ontology, biological pathways, protein domains, site of expression, cancer signatures, transcription factors, clinical phenotypes, miRNA analysis and extracellular vesicles. Other interesting features include the generation of editable venn diagrams, column/bar and pie charts as well as heatmap for analysing quantitative data.

Poster ID: 36

Vaccine-Induced Training of Innate Immunity: Challenging Immunological Dogma

Laura Sourdin (a), Damon Tumes (a), Miriam Lynn (a), Anastasia Sribnaia (a), David Williams (c), David J Lynn (a), (b).

(a) EMBL Australia Biomedical Informatics Group, Infection and Immunity Theme, South Australian Health and Medical Research Institute, Adelaide, Australia
(b) School of Medicine, Flinders University, Bedford Park, Australia
(c) Department of Surgery, James H. Quillen College of Medicine, East Tennessee State University, Johnson City, TN

Vaccines work by inducing a long-lasting, antigen-specific, memory response via the adaptive immune system, to antigens in the vaccine that are derived from a targeted pathogen. Thus, vaccine dogma is that a vaccine provides specific protection to a specific pathogen/disease but no protection to other unrelated pathogens. Emerging evidence is, however, challenging the conventional wisdom that vaccines are solely disease-specific interventions as it is becoming increasingly appreciated that vaccines can also elicit both beneficial and, more controversially, deleterious, nonspecific responses. An emerging potential mechanism through which vaccines can confer non-specific effects is through the re-programming of innate immune cells such that they respond more vigorously to subsequent unrelated antigens. This has been termed trained innate immunity. To date, very little is known regarding the molecular mechanisms which regulate trained immunity. I will present the first transcriptome-wide study which investigates the transcriptional networks that are associated with the training of human monocytes exposed in vitro to the Bacille Calmette Guérin (BCG) vaccine, a live attenuated vaccine against tuberculosis that is one of most widely administered vaccines globally, and which has been associated with significant protective non-specific effects.

Poster ID: 37

Targeted X chromosome resequencing for discovery of unknown migraine genes

Roos-Araujo, D, Sutherland, HG, Haupt, LM, Benton, MC, Lea, RA, Griffiths, LR

Genomics Research Centre, Institute of Health and Biomedical Innovation, School of Biomedical Sciences, Queensland University of Technology, Brisbane, QLD

Background: Migraine is a painful neurovascular headache disorder affecting approx. 12% of the general population. Migraine places a significant burden on the global economy due to days lost due to disability as well as the sufferers themselves. Various genes and variants have been associated with non-hemiplegic migraine susceptibility, however, a conclusive genetic cause remains unknown. An X chromosome and migraine association was discovered approx. 20 years ago and little progress has since been made in identifying an X chromosome causative variant. With this previous association and the predominance of female migraine sufferers the X chromosome remains a good candidate for further investigation.

Methodology: We performed paired end DNA-Seq on X chromosomal regions of selected X-linked migraine family pedigree members available within the Genomics Research Centre. Regions sequenced included a ~4 Mb Xq12 region and the ~37 Mb region spanning Xq24 to Xq28. Bioinformatics analyses were performed mapping sequence data to a reference genome up to variant calling and filtering.

Results: To date, our data analysis has identified a number of variants in both X chromosomal regions as potential migraine candidates and we are currently validating these in the selected individuals with Sanger sequencing. Once completed, the validated sequences will be examined in additional populations to determine their association with migraine. The identification of rare functional genetic variants associated with non-hemiplegic migraine, will provide a better understanding of the functional consequences of these variants and establish new candidate genes for future migraine research and the development of improved diagnostics tools.

Conclusions: Previous migraine genetic studies have improved our knowledge through uncovering factors contributing to migraine pathophysiology. This work aims to identify new markers of migraine diagnosis toward improving current treatments to relieve the burden of migraine on the population.

Poster ID: 38

The newly identified mutations in Streptococcus Pneumoniae serotype 3 Small Colony Variants

Yiwen Zhou (a,d), Stephen Kidd(a,b,c), Jimmy Breen(d,e), Stepthen Pederson (d)

(a) School of biological science, University of Adelaide
(b) Research Centre for infectious Disease, University of Adelaide
(c) Australian Centre for Antibiotic Resistance Ecology
(d) Bioinformatics hub, University of Adelaide
(e) Robinson Research Institute

Small colony variants (SCV) is a muti-antibiotic resistance variants evolved from Streptococcus Penumoniae with the ability of replication. Previous studies have been suggested structure variation in cps3DSU and cap3A operon leads to the formation of SCV variants. However, a global investigation in gene expression and genomic changes have not yet been thoroughly studied. Here we describe the methodology, with implementation of De novo assembly and reference-based alignments, to find out potential structural variation and single nucleotide polymorphism. Materials: Illumina Miseq 2000 Paired-end reads, Variable length 35-150 nt. Methods: For the assembly part, three assembly tools, SPades, Abyss and HGA have been put into use to achieve the best results for the reason that variable length is not always favoured in assembly tools. Apart from N50, NG50, number of contigs etc. are used as benchmarks to evaluate the best assembly results, several other evaluations are also used such as ReapR score. However, when align SCV reads align back to both Wildtype and SCV contigs, only 30% of reads are mapped perfectly. In order to complete the genome with contigs, PAGIT was implemented and newly invented assembly de Bruijn Graph REAd mapping Tool (BGREAT) to take care of repeat region. Eventually, the resulted output files (.fasta) are then proceeded to Pindel and Mauve for structural variation analysis. For alignment part, which is crucial step for variant calling, have been conducted by comparing four different alignment tools: BWA-mem, BWA-sampe, Bowtie2 and BBmap. The evaluation for those output files (.bam) are based on several criteria — mapped reads, paired reads, proper pair and mapping quality. SnpEff tools have been put into used for the SNPs callings by using selected bam file which is corrected by base alignment quality (BAQ). RESULTS: SPades assembly achieved best results in every aspects (TABLE1) with variable length paired-end reads, PAGIT significantly increased the longest contigs and decreased the the amount number of contigs. BWA-mem was selected to do variant calling based on criteria. SnpEFF and Pindel successfully found SNPs and structural variation (SV) but the those SNPS and SV is still need to be validated in wet-lab analysis.

Poster ID: 39 (COMBINE)

Exploring Pan-Cancer Network Relationships Between Somatic Changes and Expression Profiles with PACMEN

Shila Ghazanfar (a,b), Jean Yee Hwa Yang (a)

(a) The University of Sydney; (b) CSIRO

The Cancer Genome Atlas is a rich source of information enabling study across and between cancers. Recently, network approaches have been applied to such data to uncover complex interrelationships between somatic changes and expression profiles, but lack direct testing. In this pan-cancer study we co-analysed somatic changes (mutation and copy number alterations) and expression (gene or protein) to identify networks of interest, shedding light on these relationships in a direct manner.

Using somatic changes and gene expression information across each of 19 cancers, we identified mutation-expression networks and enabled interrogation through an online interactive R Shiny application, PAn Cancer Mutation Expression Networks (PACMEN). Analysis involved directly testing for differences in expression abundance via somatic changes. PACMEN allows data and parameter choice, showcases analyses performed using curated and estimated networks, and provides network description and visualisation.

We found networks identified were significantly enriched for known cancer-related genes such as melanoma ($P<0.01$ Network of Cancer Genes 4.0). Notably, comparison between cancers showed a greater overlap of nodes for cancers with higher overall mutation load (melanoma, lung), compared to those with a lower overall mutation load (glioblastoma, leukemia).

We propose and implement a framework for exploring network information through coanalysis of somatic changes and gene/protein expression profiles. Our pan-cancer approach suggests that while mutations are frequently common among cancer types, the impact they have on surrounding networks via expression changes varies, which may explain differences in efficacy of therapies among different diseases. In contrast, for some cancers mutation-associated network behaviour appears similar, suggesting a framework for uncovering cancers where similar therapeutic strategies may be applicable. PACMEN is available at <shiny.maths.usyd.edu.au/PACMEN>.

Poster ID: 40 (COMBINE)

Consequences of Drug Dose Modulation on Clonal Dynamics

Luis Lara-Gonzalez(a), David Goode (a,b), Sherene Loi (b,c), Davide Ferrari (d), Anthony Papenfuss (a,e)

(a) Bioinformatics & Cancer Genomics Laboratory, Peter MacCallum Cancer Centre, Melbourne, Victoria 3000, Australia
(b) Sir Peter MacCallum Department of Oncology, The University of Melbourne, Victoria 3010, Australia
(c) Translational Breast Cancer Laboratory, Peter MacCallum Cancer Centre
(d) Mathematical and Computational Biology, Statistics, The University of Melbourne, Victoria 3010, Australia
(e) Computational Biology, Bioinformatics Division, Walter and Eliza Hall, Victoria 3052, Australia

By reconstructing tumour evolution, computational modelling is making significant progress toward identifying drug resistance origins and optimal drug timing and order. However, data-driven assessment is lacking. We are creating an onco-bioinformatic tool to study the dynamics of tumour growth, treatment, and resistance to improve our understanding of cancer and the design of better treatments for the disease.

We modeled tumour evolution as an agent-based discrete time branching process that tracks the expansion of diverse clonal lineages as they acquire driver and passenger mutations that alter their proliferation and mutation rates. Clonal proliferation is subject to a spatio-temporal size-dependent penalty to provide characteristic tumour growth patterns. Once the tumour attains a diagnosable size (1 to 4 billion cells), a mitotic phase-specific perturbation is introduced to model anticancer agents. This environmental disruption impacts clonal dynamics, and we observed diverse heterogeneity, genomic instability, and resistance evolutionary paths. Our tool recovers various tumour development rates seen in the clinic, in which genomic instability promotes clonal diversification, leading to a state of invasiveness and prevailing (cross)-resistance.

Our tool predicts that the last couple of years prior to diagnosis are essential in the pathogenesis of the tumour, which requires 2 - 6 driver mutations to bypass the effects of anticancer agents. Our simulated clinical trials comparing cytotoxic and targeted drug combinations show that moderate-dose schemes lead to prolonged survival rates, even in the presence of pre-existing drug resistant clones. Therefore, treatments maintaining clonal proportions should be considered as an alternative way for tumour growth control.

Functional mutations form at CTCF/cohesin binding sites in melanoma due to uneven nucleotide excision repair across the motif

Rebecca C. Poulos (a), Julie A. I. Thoms (a), Yi Fang Guan (a), Ashwin Unnikrishnan (a), John E. Pimanda (a,b) and Jason W. H. Wong (a)

(a) Prince of Wales Clinical School and Lowy Cancer Research Centre, UNSW Australia, Sydney NSW 2052, Australia
(b) Department of Haematology, Prince of Wales Hospital, Sydney NSW 2052, Australia

CCCTC-binding factor (CTCF) is a crucial protein involved in maintaining the three-dimensional organisation of the genome and defining regions of gene expression. CTCF binding sites are frequently mutated in cancer, but how these mutations accumulate and whether they broadly perturb CTCF binding is not well understood. In this study, we analysed the genomes of 52 skin cancer samples and found skin cancers to exhibit a unique and asymmetric mutation pattern within CTCF motifs. To understand the mechanisms underlying mutation formation, we examined datasets of nucleotide excision repair (NER), CTCF binding and replication timing, showing the mutation pattern to be attributable to ultraviolet irradiation and differential NER across individual nucleotides within CTCF motifs. We additionally demonstrated that CTCF binding site mutations form independent of replication timing, with mutations enriched at sites of CTCF/cohesin complex binding and suggesting a novel role for cohesin in stabilizing CTCF-DNA binding and impairing NER in a multifactorial manner. To demonstrate that CTCF binding site mutations are functional in melanoma, we performed CTCF ChIP-seq in a melanoma cell-line, reporting allele-specific reduction of CTCF binding to mutant alleles. Investigating whether selection underlies CTCF motif mutation accumulation, we analysed topologically-associated domains in skin cells and identified mutated CTCF anchors which contain differentially-expressed cancer-associated genes. One gene identified was the tumour suppressor APC, a key component of the Wnt signaling transduction pathway and a gene previously implicated in melanoma development. Through bootstrapping analyses and mutation simulation however, we found that genome-wide, CTCF motif mutations in melanoma are generally under neutral selection. Regardless, the frequency and potential functional impact of such mutations highlights the need to consider their impact on cellular phenotype in personalized genomes.

Poster ID: 42 (COMBINE)

GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly

Daniel L Cameron (a,b), Anthony T Papenfuss (a,b,c)

(a) Bioinformatics Division, Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria, 3052, Australia
(b) Department of Medical Biology, University of Melbourne, Parkville, Victoria, 3010, Australia
(c) Peter MacCallum Cancer Centre, Victorian Comprehensive Cancer Centre, Melbourne, 3000, Australia

The identification of genomic rearrangements with high sensitivity and specificity using massively parallel sequencing remains a major challenge. Many methods have been developed for Illumina sequence data, with most methods using read depth analysis, read pair clustering, split read identification, assembly, or a combination of these approaches. Existing assembly-based methods perform either de novo assembly (e.g. cortex), targeted assembly based on previously identified candidates (e.g. manta, SVMerge, TIGRA), or perform windowed assembly to detect small events (e.g. DISCOVAR, SOAPindel).

Here we describe GRIDSS, the Genome Rearrangement IDentification Software Suite, composed of an assembler, and a variant caller which combines assembly, split read and read pair evidence to identify genomic rearrangement breakpoints using a probabilistic model. Our novel genome-wide break-end assembly approach assembles reads not supporting the reference prior to breakpoint identification or variant calling using a positional de Bruijn graph. By constraining the assembly of each read based on the mapping locations of the read/read pair, and encoding these assembly constraints directly within the assembly graph itself, a single genome-wide assembly can be performed.

GRIDSS recently won structural variant detection sub-challenge #5 of the ICGA-TCGA DREAM Somatic Mutation Calling challenge and has been extensively benchmarked against BreakDancer, cortex, CREST, DELLY, HYDRA, LUMPY, manta, Pindel, SOCRATES and TIGRA across a wide range of simulated variant types, variant sizes, read depths, read lengths, and library fragment sizes. With the exceptions of low coverage data (≤8x), and large novel insertions detectable only by de novo assemblers, GRIDSS F-scores exceeded that of all other callers. On well-studied human cell line data, GRIDSS is able to achieve a false discovery rate less than half that of other methods, with no loss of sensitivity.

Poster ID: 43 (COMBINE)

How do vulnerabilities left during the evolution of cellular networks set the stage for cancer?

Trigos, A.S. (a,b), Pearson, R.B. (b,c,d), Papenfuss, A.T. (a,b,e), Goode, D.L. (a,b)

(a) Bioinformatics and Cancer Genomics Laboratory, Peter MacCallum Cancer Centre, Melbourne, Victoria 3000, Australia.
(b) Sir Peter MacCallum Department of Oncology, The University of Melbourne, Victoria 3010, Australia.
(c) Department of Biochemistry and Molecular Biology, The University of Melbourne, Parkville, Victoria 3010, Australia.
(d) Department of Biochemistry and Molecular Biology, Monash University, Clayton, Victoria 3168, Australia.
(e) Bioinformatics Division, The Walter & Eliza Hall Institute of Medical Research, Parkville, Victoria 3052, Australia.

Taking a top-down approach from protein-protein and gene-regulatory networks to derive biologically relevant information and key genes is challenging given their complexity. To bridge the gap between network-level and gene-level analysis, we tackled an intermediary level of organization, interactions between biological processes.

We developed a metric for the interconnectedness of high-level cellular processes (GOslims) based on the entire human network of gene and protein interactions to construct a network of co-regulation. This metric normalizes the number of gene-gene edges connecting pairs of processes by the total possible number of edges. Overlaying co-expression data from 7 tissue types onto our co-regulation network identified patterns of transcriptional and functional dependency.

A strong positive correlation of expression of highly interconnected processes involved in basal cellular functions suggests co-regulation developed by a long history of co-evolution. In contrast, consistent negative correlation in expression indicates incompatibility between the activation of basal and more recently evolved cellular functions, suggesting strong mutual exclusivity.

Expression data from The Cancer Genome Atlas demonstrated how these patterns are altered in cancer. We detected an enhanced mutual exclusivity of unicellular and multicellular functions, with many pairs of processes key to carcinogenesis positively correlated in normals becoming negatively correlated in tumours. Extending our approach to the gene level, we found the same genes modulated these alterations across multiple tumours, uncovering putative novel common drivers.

Our results indicate the co-regulatory networks between cellular processes can modulate adaptive and evolutionary processes, and their disruption can drive tumour progression. The study of these networks provides a level of organization that facilitates the link between network-level and gene-level analyses.

Poster ID: 44 (COMBINE)

Assessing the Practicality of Oxford Nanopore Sequencing in Clinical Diagnositcs

Alexis Lucattini (a,b), Lavinia Gordon (a), Matt Ritchie (b)

(a) Australian Genome Research Facility
(b) Walter & Eliza Hall Institute for Medial Research

Much excitement has arisen over the new Oxford Nanopore's MinION and it's potential in real-time diagnostic sequencing applications. Unfortunately in order to obtain high-molecular DNA many samples provide insufficient DNA to extract and sequence directly, requiring overnight growth following by extraction. Thus hindering the real-time advantage of the platform. The new Oxford Nanopore R9 kit comes with the potential of 'rapid sequencing', requiring only a ten minute library-preparation. Using the QIAGEN-Mini Kit we were able to obtain a sufficient DNA extraction within 120 minutes followed by the rapid 1D preparation. With local base calling, we compared the output of using a local metagenomic database; OneCodex; and WIMP (What's in my pot?) Oxford Nanopore's metagenomic classifier. We found that the OneCodex database is the most extensive of the three, with a much more sensitive aligner. Our own local database allowed for fast identification but only classified 5% of the reads. WIMP had a similar classification level to our own local database but lagged due to upload/download time. We present all three using Krona, which displays interactive hierarchical data on zoomable piecharts.

Poster ID: 45 (COMBINE)

Exploiting extremes of Legionella pneumophila genomic diversity for accurate source attribution

Andrew H. Buultjens (a), (b), Kyra Y. L. Chua (c), Sarah L. Baines (a), Jason Kwong (a), (b), (c), Wei Gao (b), Mark B. Schultz (a), (c), Zoe Cutcher (d), (e), Stuart Adcock (d), Susan Ballard (a), Takehiro Tomita (a), Nela Subasinghe (a), Glen Carter (b), (c), Sacha J. Pidot (b), (c), Lucinda Franklin (d), Torsten Seemann (c), (f), Anders Gonçalves Da Silva (a), (c), Benjamin P. Howden (a), (b), (c), Timothy P. Stinear (b), (c)

(a) Department of Microbiology and Immunology at the Peter Doherty Institute for Infection and Immunity, The University of Melbourne, Victoria, Australia
(b) Doherty Applied Microbial Genomics, The Peter Doherty Institute for Infection and Immunity, Victoria, Australia
(c) Microbiological Diagnostic Unit Public Health Laboratory at the Peter Doherty Institute for Infection and Immunity, The University of Melbourne, Victoria, Australia.
(d) Health Protection Branch, Department of Health and Human Services, Victoria, Australia
(e) National Centre for Epidemiology and Population Health, Australian National University, Canberra, Australia
(f) Victorian Life Sciences Computational Initiative, The University of Melbourne, Victoria, Australia

Public health agencies are increasingly using genomics in Legionnaires' disease investigations, relying on similarities in the core genomes of the causative bacteria (Legionella pneumophila) to identify environmental contamination sources. Here, we show that the assumptions underlying these studies are flawed. We propose instead a statistical learning approach for source attribution that also considers accessory genome variation. We compared genomes of 234 L. pneumophila isolates obtained from patients and cooling towers in Melbourne, Australia between 1994 and 2014. This collection spanned 29 infection clusters, including one of the largest reported Legionnaires' disease outbreaks, involving 125 cases at an aquarium. There was one dominant genotype that exhibited startlingly low core genome variation. The median pairwise nucleotide difference for the 180 genomes obtained across Melbourne over 21 years was only 5 single nucleotide polymorphisms (SNPs) (IQR 3-7). In addition to high sequence conservation, we also uncovered within-outbreak isolate diversity. By assessing only cooling tower isolates and including all genomic variation (SNPs and accessory genome), we built a multivariate model to find cooling tower-specific genomic signatures. We then used this model to accurately predict the origin of clinical isolates. A sister model built with SNPs from the recently advocated cgMLST approach applied to this same population was poorly predictive. These data show that health agencies will require a deep understanding of local L. pneumophila population structure or risk source misattribution. The restricted genetic diversity seen here also suggests environmental reservoirs of quiescent bacteria sporadically seeding warm water sources, causing human cases of Legionellosis.

Poster ID: 46 (COMBINE)

Tracking clonal evolution in cancer from multiple samples

Christoffer Flensburg, Ian Majewski

WEHI, Melbourne, Australia

Cancer is constantly evolving. To understand the disease, we need to monitor how it changes, and the advent of genome wide DNA and RNA sequencing provides a powerful way of doing this. Our lab has developed methods to take full advantage of sequencing data from multiple cancer samples from a single individual. This allows us to track the clonal evolution of a cancer: pinpointing molecular changes in cancer cells that resist therapy, spread around the body or transform into more aggressive diseases. I will illustrate some of these methods through a range of examples.

We identify clonal populations based on somatic mutations, both single nucleotide variants and copy number alterations. In some individuals we see genes or pathways being repeatedly mutated in distinct cell population, which can identify driver mutations from small cohorts, or even from a single individual. We see the same copy number alteration present in multiple samples affect different alleles. This shows that the copy number alterations in the two samples are different events, which again points towards a driver mutation. Clonal consistency constraints allow us to filter noise and inform the choice of phylogenetic tree of the cancer. We combine captured DNA and RNA from the same sample to identify cis acting regulatory mutations that could not have been found from the DNA or RNA alone.

In combination, these methods enable significant findings from multiple cancer samples that would not be possible from a naïve cancer-normal analysis.

Poster ID: 47 (COMBINE)

BootNet: a bootstrapping application for GLMnet modelling to identify robust classifiers in genomics data

Rod Lea (a), Nicole White (b), Ray Blick (c), Macartney-Coxson (c), Daniel Kennedy (b), Lyn Griffiths (a), and Miles Benton (a)

(a) Genomics Research Centre, Institute of Health and Biomedical Innovation, Queensland University of Technology, Brisbane, Queensland, Australia
(b) ARC Centre of Excellence for Mathematical and Statistical Frontiers, Queensland University of Technology (QUT), Brisbane, Queensland, Australia
(c) Kenepuru Science Centre, Institute of Environmental Science and Research, Wellington, New Zealand

The R-based GLMnet package is a powerful machine learning classifier that provides benefits to analysis of highly dimensional genomic data through the implementation of an elastic-net routine. This framework brings together two established approaches, penalised ridge regression and LASSO, and by applying specific tuning parameters is able to overcome limitations conventional methods such as OLS regression. We developed a wrapper to GLMNet (BootNet) that adds fast bootstrap resampling algorithms to all types of outcome and predictor variables. BootNet accepts both quantitative and qualitative outcomes, i.e. tissue type, disease state, blood pressure, age. The base algorithm employs a percentage-based sub-sampling routine. We have also implemented a Jackknife approach (for outlier detection) as well as a leave-one-out cross-validation method. Additionally, we have added the ability to run processing in parallel, greatly reducing the time of computation on both local machines and servers. In testing we have explored numerous methylation data sets from various tissues with great success. Using a discovery cohort we were able to select CpG sites that differentiate abdominal from omental adipose 100% of the time. In another experiment we demonstrated that the BootNet method could identify robust DMRs and single CpGs associated with aging in a healthy cohort, with potential biological relevance and statistical significance. We were able to validate our top findings in the large cohort (n=2316) of publicly available methylation data from MARMAL-AID. The elastic-net methods implemented in GLMnet are very powerful for classifying data, and with our BootNet wrapper this package now has potential to be developed into a fast integrative ‘omics approach for identification of robust classifiers of states such as disease outcomes.

Poster ID: 48 (COMBINE)

Estimating genetic similarity with the k-mer Weighted Inner Product (kWIP)

Kevin Murray (a), Cheng Soon Ong (b, c), Christfried Webers (b,c), Justin Borevitz (a), Norman Warthman (a)

(a) Centre of Excellence in Plant Energy Biology, The Australian National University, Canberra, Australia
(b) Data61, CSIRO, Canberra, Australia
(c) Department of Computer Science, The Australian National University, Canberra, Australia

Modern genomics techniques generate overwhelming quantities of data. Extracting population genetic variation demands computationally efficient methods to determine genetic relatedness between individuals or samples in an unbiased manner, preferably de novo. The rapid and unbiased estimation of genetic relatedness has the potential to overcome reference genome bias, to detect mix-ups early, and to verify that biological replicates belong to the same genetic lineage before conclusions are drawn using mislabeled, or misidentified samples.

We present the k-mer Weighted Inner Product (kWIP), an assembly-, and alignment-free estimator of genetic similarity. kWIP combines a probabilistic data structure with a novel metric, the weighted inner product (WIP), to efficiently calculate pairwise similarity between sequencing runs from their k-mer counts. It produces a distance matrix, which can then be further analysed and visualized. Our method does not require prior knowledge of the underlying genomes and applications include detecting sample identity and mix-up, non-obvious genomic variation, and population structure.

We show that kWIP can reconstruct the true relatedness between samples from simulated populations. By re-analyzing several published datasets we show that our results are consistent with marker-based analyses. kWIP is written in C++, licensed under the GNU GPL, and is available from https://github.com/kdmurray91/kwip.

Poster ID: 39 (AB3ACBS)

Insights into the signalling pathways studies of GPR139 and impacts of their inhibitors using systems biology and pharmacokinetics

Aman Chandra Kaushik (a, b), Deeksha Gautam (a), Shakti Sahi(a)

(a) School of Biotechnology, Gautam Buddha University, Greater Noida, Uttar Pradesh, India
(b) The Shraga Segal Dept. of Microbiology, Immunology and Genetics, Faculty of Health Sciences, Ben-Gurion University of the Negev, Beer-Sheva, Israel

GPCR share a common structural feature i.e. the presence of seven membrane spanning helices sharing three each intracellular and extracellular loops with an intracellular carboxyl terminal and an extracellular amino terminal. Various differences in structure of GPCR can occur during the ligand binding, G protein coupling and interaction with other proteins. In Rhodopsin class of GPCR the basic structure of GPCR is verified by X-ray crystallography. Ligands acts as an extracellular stimulus for GPCR's and then brings physiological changes in organisms. In this article construct biochemical pathway of Alzheimer, Parkinson and Type 2 diabetes and associated with GPR139. GPR139 found most effective physiological role in all peripheral organs and as well as in central nervous system. GPR139 targeted by virtually screened compounds which inhibit the Alzheimer, Parkinson and Type 2 diabetes. Molecular Dynamics simulation was performed for complex structure of GPR139 to validate binding affinity of compounds, and investigate active site fluctuation of complex structure.

Poster ID: 40 (AB3ACBS)

Establishing Cell Composition in Microarray and RNA-seq Data

Saskia Freytag (a), Johann Gagnon-Bartsch (b), Terry Speed (c) and Melanie Bahlo (a)

(a) Population Health and Immunity Division, The Walter and Eliza Hall Institute of Medical Research
(b) Michigan Institute of Data Science, University of Michigan
(c) Bioinformatics Division, The Walter & Eliza Hall Institute of Medical Research

Both microarray and RNA-seq technology are frequently used to measure gene expression in samples consisting of many different cell types. However, gene expression differs between cell types to create the specialized functions and physiologies of many different cell types. Differences in gene expression due to cell type have been demonstrated to drive differential expression results in a recent study of psychiatric disorders. Given that cell composition can also change as a result of diseases, such as Huntington's disease or schizophrenia, it is important to account for differences in cell-type proportions in differential expression analysis. Since sample compositions are rarely measured, they have to be accurately inferred using computational algorithms, referred to as deconvolution.

We recently developed a new simple deconvolution approach based on ratios. We demonstrated that this new approach is able to infer cell composition more accurately and robustly than any of the existing approaches in a variety of test datasets with true cell proportions established either by fluorescence-activated cell sorting or by experimental design. Moreover, our approach is fast, scalable and easy to optimize. Finally, we showed that it can be applied to real data from post-mortem human brain, which consists if one of the most diverse collections of cell types.

Poster ID: 41 (AB3ACBS)

Consensus modelling reveals systematic differences in precision and sensitivity between platforms measuring DNA methylation

Timothy J Peters(a), Terence P Speed(b), Ruth Pidsley(a), Elena Zotenko(a) and Susan J Clark(a)

(a) Genomics and Epigenetics Division, Garvan Institute of Medical Research, Darlinghurst, NSW 2010, Australia
(b) Walter and Eliza Hall Institute, Parkville, Victoria 3052, Australia

Introduction: The study of DNA methylation in the human genome has benefited recently from the development of technologies such as whole genome bisulphite sequencing (WGBS) and Illumina Infinium BeadChip arrays. These platforms allow interrogation of methylation levels at single nucleotide resolution, allowing granular insights and opportunities for validation of original findings. However, care must be taken when denoting a particular platform as a "gold standard", since this immediately introduces a bias.

Methods: Since all measurements of DNA methylation are estimates, in the absence of a "gold standard" we instead assess the precision and sensitivity of three different methylation platforms via a consensus modelling method: two Illumina array platforms and WGBS run on the same suite of samples. This method was developed and used by the United States Bureau of Standards to assess interlaboratory precision and sources of variability in manufacturing the same product across multiple sites.

Results: We find that array platforms are prone to decreased sensitivity to methylation change in a subset of probes, potentially due to elements of probe design impeding hybridisation specificity. There is also evidence for a relationship between the depth of coverage of the WGBS and the precision of the methylation measurement.

Conclusion: Our findings demonstrate limits to the degree of concordance between measurements on these platforms. We recommend taking into consideration probe biochemistry and sequencing depth when undertaking validation studies using these platforms.

Poster ID: 42 (AB3ACBS)

Koala retrovirus (KoRV) insertion behaviours within the Koala genome.

R. Salinas (a), Z. Chen (a,b), M. Hobbs (c), A. King (d), P. Timms (e), M. Wilkins (a) and the Koala Genome Consortium

(a) School of Biotechnology and Biomolecular Sciences, UNSW Australia Sydney NSW 2052 Australia
(b) Illumina Australia and New Zealand, 1 International Court, Scoresby, Victoria, 3179
(c) Garvan Institute of Medical Research, 384 Victoria Street, Darlinghurst Sydney NSW 2010 Australia
(d) Australian Museum, 1 William St, Sydney NSW 2010
(e) Faculty of Science, Health, Education and Engineering, University of Sunshine Coast, 90 Sippy Downs Dr, Sippy Downs QLD 4556

The koala (Phascolarctos cinereus) is an Australian arboreal marsupial that has been found to have widespread infection with the Koala retrovirus (KoRV) within its populations. Although previous studies have isolated the retrovirus and sequenced its retroviral genome, little work has been done to characterise its interactions with the host organism after the retrovirus' genomic insertion.

The initial sequencing of the koala genome by the Koala Genome Consortium (KGC) was done with the Illumina HiSeq 2500 platform and this was used to produce KoRV targeted assemblies. In this work, we have characterised the insertion of the KoRV within the host genome . The KoRV inserts were isolated and motif analysis of the insert point within the genome was analysed for preference of position of insertion in the koala genome and the specificity of KoRV components. This analysis produced a repetitive genomic koala motif and a KoRV motif corresponding to its LTR.

Later whole genome sequencing of koala was done by KGC using the Pacific Biosciences RSII platform. We have used the PacBio reads to perform a quantitative analysis of the KoRV insertions to specifically verify the inserts' sizes and validate the Illumina-based motifs, resulting in the observed truncation of KoRV inserts.

Poster ID: 43 (AB3ACBS)

The Network Basis of Negative Genetic Interactions in Saccharomyces cerevisiae

Chi Nam Ignatius Pang, Apurv Goel and Marc R. Wilkins

Systems Biology Initiative, School of Biotechnology and Biomolecular Sciences, University of New South Wales, NSW 2052, Australia.

The basis of negative genetic interactions is poorly understood. To investigate this, we have integrated biological networks from Saccharomyces cerevisiae, combining high confidence protein-protein interactions, kinase-substrate interactions and transcription factor-target gene relationships. Triplet motifs in the networks, each consisted of one pair of proteins that show negative genetic interaction, and a third interacting protein, were enumerated and analyzed. Negative genetic interactions do not arise at random, but are significantly over-represented in 7 out of the 14 types of triplet motifs present. These 7 types of motifs were associated with multiprotein complexes and signaling pathways, including in feed-forward loops. In addition, we observed an enrichment of essential third proteins, which interacts with the negatively interacting genetic pair, among 5 types of motif. The majority of negative interactions were shared among triplet motifs that are over-represented in the networks, but are not explained by the co-deletion of orthologs that abrogate genetic redundancy. We also noted that negative genetic interactions are frequently observed when the proteins are periodically co-expressed, such as in the cell cycle. Together, our results highlight that negative genetic interactions are significantly associated with important roles in networks and help define the fundamental logical processing units in the eukaryotic cell.

Poster ID: 44 (AB3ACBS)

Deciphering the role of Campylobacter concisus in host pathogenesis using multi-omics techniques

Nandan P. Deshpande (a), Marc R. Wilkins(a,b), Hazel Mitchell (b), Nadeem O. Kaakoush (c)

(a) Systems Biology Initiative, UNSW
(b) School of Biotechnology and Biomolecular Sciences
(c) School of Medical Sciences, UNSW Australia, Sydney, New South Wales, Australia

Campylobacter concisus, a commensal within the oral cavity, is an emergent opportunistic pathogen of the gastrointestinal tract associated with a range of diseases. In order to understand the virulence make-up of C. concisus, we sequenced and assembled several isolates with different motility, adherence and invasion capabilities followed by comprehensive comparative genomics analysis. This revealed high genetic diversity across the strains, and identified both genes and non-synonymous SNPs associated with the multiple phenotypes. For example, strains with an ability to survive intracellularly within host epithelial cells had distinct SNPs in genes within the peptidoglycan biosynthesis pathway, and an additional extrachromosomal restriction-modification system. In contrast, a subset of strains had within their genome a zonula occludens toxin, known in other bacteria to target host tight junctions. These findings led to the hypothesis that C. concisus strains belong to potential pathotypes such as adherent-invasive, adherent-toxigenic and non-pathogenic C. concisus strains. Given this, the host epithelial and immune responses to the different C. concisus pathotypes was assessed using techniques including RNA-Seq, mass spectrometry and qPCR. Comprehensive global profiles of the responses to C. concisus infection were determined. Of interest, nucleic acid sensing was shown to be important for host recognition of this bacterium.

Poster ID: 45 (AB3ACBS)

Enrichment of PRC2-binding RNA motifs for RNA-mediation regulation

Haroon Naeem(a,b), Stuart Archer(a,b), David Powell(a,b) and Chen Davidovich(b,c)

(a) Monash Bioinformatics Platform, Monash University, Victoria 3800, Australia.
(b) Biomedicine Discovery Institute, Faculty of Medicine, Nursing and Health Sciences, Monash University, Victoria 3800, Australia.
(c) Department of Biochemistry and Molecular Biology, Monash University, Victoria 3800, Australia; EMBL Australia and the ARC Centre of Excellence in Advanced Molecular Imaging, Clayton, VIC 3800, Australia.

Motivation: Polycomb repressive complex 2 (PRC2) has been discovered as an important chromatin regulator of epigenetic gene silencing. PRC2 is interacting with thousands of nascent RNA transcripts that has been proposed to regulate its primary function as an H3K27 histone methyltransferase. However, little is known about how PRC2 identifies these RNAs.

Results: We empirically quantified the genome-wide and transcriptome-wide association between PRC2 and G-tract rich motifs of the sequences [GnNm]k, where G is a guanine and N can be any nucleotide. We show that motifs composed of multiple short G-tract repeats are significantly enriched at PRC2 binding sites within transcripts. Importantly, analysis of ChIP-Seq data shows that DNA sequences coding for PRC2-binding RNA motifs are enriched at PRC2 binding sites on chromatin and significantly associate with H3K27 tri-methylation; the product of PRC2, exclusively.

Conclusion: The significant enrichment of PRC2-binding RNA motifs at polycomb target transcripts provides means for RNA-mediated epigenetic regulation.

Poster ID: 46 (AB3ACBS)

Navigating the Research Data Life Cycle

Philippa Griffin (a), Rudi Appels (b,c), Dieter Bulach (a), Kevin Dudley (d), Gabriel Keeble-Gagnere (b), Andrew Pask (e), Bernard Pope (a) , Ute Roessner (e), Torsten Seemann (a), Dan Bolser (f), Jyoti Khadake (g), Suzanna Lewis (h), Sandra Orchard (f),

(a)EMBL-ABR: VLSCI node, Victorian Life Sciences Computation Initiative, Melbourne, Australia
(b)AgriBio, Centre for AgriBioscience, Melbourne, Australia
(c)Murdoch University, Perth, Australia
(d)Queensland University of Technology, Brisbane, Australia
(e)The University of Melbourne, Melbourne, Australia
(f)European Bioinformatics Institute, Hinxton, UK
(g) NIHR BioResource, University of Cambridge, Cambridge, UL
(h) Berkeley Bioinformatics Open-source Project, Lawrence Berkeley National Laboratory, Berkeley, USA
(i) EMBL-ABR: AGRF Node, Australian Genome Research Facility, Melbourne, Australia
(j) EMBL-ABR HUB, Melbourne, Australia

Where does your research data go once you've published your paper? Can you do better? Good data management spans all stages of the data life cycle: finding, collecting, integrating, processing, visualising, analysing, publishing, sharing and reusing data and metadata. EMBL Australia Bioinformatics Resource (EMBL-ABR) aims to increase Australia's capacity to deal with the large heterogeneous data sets now part of modern life science and biomedical research, in line with FAIR principles, so that data are Findable, Accessible, Interoperable and Reusable. In this poster we present recent efforts and international engagement in data life cycle best-practice for Australian life sciences such as agri-food research, animal research, biomedical research and big data for science. We will also include results from our best practice workshop series on data life cycle on: (i) mapping the current state of the life-cycle in the Australian biological domain and bioinformatics, (ii) highlighting the existence of relevant Australian research data sets to the international environment and (iii) building the roadmap for the bioinformatics resources and tools with those working in the specific biological domains in Australia.

Poster ID: 47 (AB3ACBS)

Genetically Perturbed Pathways and Core Driver Genes in Late-Onset Alzheimer's Disease

Song Gao (a), Aaron Casey (a), Tim Sargeant (d), Ville-Petteri Mäkinen (a,b,c)

(a) Heart Health, South Australian Health & Medical Research Institute, Adelaide, South Australia
(b) School of Biological Sciences, University of Adelaide, Adelaide, South Australia
(c) Computational Medicine, Faculty of Medicine, University of Oulu and Biocenter Oulu, Oulu, Finland
(d) Lysosomal Diseases Research Unit, Nutrition and Metabolism, South Australian Health & Medical Research Institute, Adelaide, South Australia

Late-onset Alzheimer's disease (LOAD) is the most common cause for dementia in the elderly, yet there are no effective therapies available that would halt the neurodegenerative process. Genetic studies have implicated APOE alleles as significant modifiers of LOAD risk, and age-related decline in lysosomal function may explain the histological hallmarks of LOAD, but the disease mechanisms are not fully understood. To bridge this knowledge gap, we integrated genetic LOAD data with public databases using the Mergeomics pipeline. GWAS of 5,896 African American participants and 54,162 participants with European ancestry, and exome sequencing data of 10,914 individuals were used for LOAD analysis. In addition, a Parkinson's disease GWAS of 108,990 individuals was included. A previously published lysosome gene set and curated pathways from MSigDB v5.0 (Broad Institute, USA) and drug signatures from DSigDB (University of Colorado, USA) were tested for aggregate genetic signals. Protein-protein interaction network was obtained from STRING v10 (STRING Consortium, Europe). The lysosome gene set was perturbed both in LOAD ($P = 2.30 \times 10{-}6$) and Parkinson's disease ($P = 2.00 \times 10{-}4$). Surprisingly, the LOAD lysosome signal was not explained by APOE alone ($P = 9.71 \times 10{-}5$ after removing APOE loci). Immune related pathways from MSigDB were also implicated in LOAD ($P = 5.25 \times 10{-}6$), but none of the drug signatures from DSigDB were statistically significant. Network analysis indicated that the genetic signals for LOAD contained several genes, such as MMP14, SREBF2 and TREM2, some of which have been linked to LOAD previously. APOE was also among the top key driver signals within the lysosome gene set. These results confirm the polygenic nature of LOAD beyond APOE, and support the hypothesis that perturbations to lysosomal function may predispose to common neuro-degenerative diseases.

Poster ID: 48 (AB3ACBS)

Prescreening biomarkers for clinical trials: feature selection and model prediction using four well known statistical methods

James Doecke

CSIRO Health and Biosecurity/ Australian E-Health Research Centre

With an ever increasing need to define small sets of predictive biomarkers for clinical trial pre-screening, many different statistical methods are used for feature selection and model prediction. Due to variable assay and collection methods, noisy data often aligns with inconsistent biomarker selection, resulting in poor validation across cohorts. In this work, we compare four common statistical methods in their ability to A) select an optimum set of features, and B) to predict outcome using biomarker sets. Feature selection methods included least least absolute shrinkage and selection operator, generalized boosted trees, random forest and stepwise logistic regression. Optimal statistical models were tested using the predicted values from the training model with the unseen test data using the receiver operating characteristic test. Predictive capability (via the area under the curve (AUC)) was assessed across increasing numbers of biomarkers added to the model. AUC values were compared from biomarker sets across biomarker number and statistical method. Real data was used from the Australian Imaging, Biomarkers and Lifestyle study, with ~1,000 features/samples. Seventy percent of the data was selected for training with the remaining 30% set aside for testing. Within the training data, random selections were portioned for iterative processing with each statistical method. Results show variable sets of biomarkers selected between different methodologies, with only a small homogenous fraction. Optimum biomarker sets ranged between 2 and 15 markers, with a positive relationship between AUC (0.64-0.86) and biomarker number, in all but the GLM. In summary, we defined optimal sets of biomarkers using four different statistical models, with only a small number of biomarkers selected across all methodologies. Strict statistical methodology and pre-processing of raw data is more likely to define biomarkers that pass external validation.

Poster ID: 49 (AB3ACBS)

Combining multiple tools outperforms individual methods in gene set enrichment analyses

Monther Alhamdoosh(a), Milica Ng(a), Nicholas J. Wilson(a), Julie M. Sheridan(b,c), Huy Huynh(a), Michael J. Wilson(a), Matthew E. Ritchie(d,e)

(a) CSL Limited, Bio21 Institute, 30 Flemington Road, Parkville, Victoria 3010, Australia.
(b) ACRF Stem Cells and Cancer Division, The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, Victoria 3052, Australia.
(c) Department of Medical Biology, The University of Melbourne, Parkville, Victoria 3010, Australia.
(d) Molecular Medicine Division, The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, Victoria 3052, Australia.
(e) School of Mathematics and Statistics, The University of Melbourne, Parkville, Victoria 3010, Australia.

Gene set enrichment (GSE) analysis allows researchers to efficiently extract biological insight from long lists of differentially expressed genes by interrogating them at a systems level. In recent years, there has been a proliferation of GSE analysis methods and hence it has become increasingly difficult for researchers to select an optimal GSE tool based on their particular data set. Moreover, the majority of GSE analysis methods do not allow researchers to simultaneously compare gene set level results between multiple experimental conditions. The ensemble of genes set enrichment analyses (EGSEA) is a method developed for RNA-sequencing data that combines results from twelve algorithms and calculates collective gene set scores to improve the biological relevance of the highest ranked gene sets. EGSEA's gene set database contains around 25,000 gene sets from sixteen collections. It has multiple visualization capabilities that allow researchers to view gene sets at various levels of granularity. EGSEA has been tested on simulated data and on a number of human and mouse data sets and, based on biologists' feedback, consistently outperforms the individual tools that have been combined. Our evaluation demonstrates the superiority of the ensemble approach for GSE analysis, and its utility to effectively and efficiently extrapolate biological functions and potential involvement in disease processes from lists of differentially regulated genes.

Poster ID: 50 (AB3ACBS)

rnaCleanR: A tool for quantifying and removing DNA contamination from strand-specific RNA-seq

Thu-Hien To and Steve Pederson

Bioinformatics Hub - The University of Adelaide

Abstract: Strand-specific RNA-seq preserves information about the strand of origin for the original mRNA template molecule. . However, there is always a certain quantity of DNA contamination present in an RNA-seq experiment, which can effect the downstream analysis such as detection of differentially expressed genes. We propose here an approach to quantify and remove DNA contamination using the strand-specific information of the RNA library. This approach uses a sliding window along the genome, and estimates the probability that reads within each window derive from stranded RNA. Reads within a window are then removed or retained based on this probability. The effects on differential expression due to varying levels of DNA contamination, and subsequent cleaning were then assessed by simulation. The results show that DNA contamination can significantly impede detection of differentially expressed genes, and depending on the initial contamination level, our approach can rescue up to 70% of lost differentially expressed genes.

Poster ID: 51 (AB3ACBS)

igLibQC: Quality Control Pipeline for the Construction of Antibody Libraries

Monther Alhamdoosh, Chao-Guang Chen, Milica Ng, Michael Wilson, Con Panousis

CSL Limited, Bio21 Institute, 30 Flemington Road, Parkville, Victoria 3010, Australia.

During the past two decades, a large number of antibody libraries have been constructed to meet the needs of drug discovery and diagnostic processes. The advent of next-generation sequencing (NGS) technology has enabled scientists to rigorously assess library size, quality, diversity and robustness at different stages of the construction process. The currently available bioinformatic tools mainly focus on the analysis of clonotypes of T cell receptors. We propose a new software pipeline, igLibQC, designed to facilitate a high-throughput analysis of NGS reads of the variable (V) domain of an antibody chain. The igLibQC pipeline includes all the essential analysis steps from merging paired-end reads, annotating V-(D)-J rearrangement, estimating the abundance levels of germline genes and families, visualising the alignment quality of germline genes (including the filtering of low quality sequences), predicting frame shifts and identifying functional clones, and finally calculating spectratypes and clonotypes to estimate the diversity of the library. Importantly, igLibQC also contains functionality to facilitate the selection of the best combination of restriction enzymes for the construction of library vectors. We illustrate the pipeline capabilities by applying it to a naïve IgM repertoire extracted from peripheral blood lymphocyte (PBL) of pooled human donors. The results show that the abundance of germline genes is inline with the natural distribution that is reported in the literature. The integrity of frameworks, complementarity-determining regions and secretion signals has been examined through comprehensive motif analysis. Overall, the results confirm that the repertoire is not pathological and can be used for library construction.

Poster ID: 52 (AB3ACBS)

Lace: constructing a transcriptome reference

Anthony Hawkins (a), Alicia Oshlack (a,b), Nadia davidson (a)

(a) Murdoch Childrens Research Institute, Royal Children's Hospital, Victoria, Australia
(b) School of BioSciences, University of Melbourne, Victoria, Australia

Transcriptomes are complex, diverse and dynamic. RNA sequencing has the potential to elucidate the richness and dynamism of gene expression, alternative splicing and ultimately underlying biological processes. However there are still limitations in analysis methods for exploring the complexity of the data. In particular, the visualization and analysis of alternative isoform expression is difficult. In order to address this we present Lace, a python package to construct superTranscripts. A superTranscript is a single linear sequence representing each gene that contains all the unique exonic sequence from all transcripts whilst retaining transcriptional ordering. Lace enables superTranscripts to be constructed by combining transcripts from any source: including de novo assembly and reference transcripts. Lace uses an overlap approach to construct a directed graph, which is then topologically sorted to produce the superTranscript. We show how superTranscripts, constructed with Lace, enable novel visualization and analysis of transcriptomes from both model and non-model organisms.

Poster ID: 53 (AB3ACBS)

Wheat Annotation Platform

Rudi Appels(a,b), Gabriel Keeble-Gagnere (a), Ute Roessner (c), Madison Flannery (d), Simon Gladman (d), Sonika Tyagi (e), Andrew Lonie (d), Antony (Tony) Bacic (c) and Maria Victoria Schneider (e)

(a)AgriBio, Centre for AgriBioscience, Melbourne, Australia
(b)Murdoch University, Perth, Australia
(c)School of BioSciences, The University of Melbourne, Australia
(d)EMBL-ABR: VLSCI node, Victorian Life Sciences Computation Initiative, Melbourne, Australia
(e)EMBL-ABR: AGRF Node, Australian Genome Research Facility, Melbourne, Australia
(f)EMBL-ABR Hub, Melbourne, Australia

The wheat genome is extremely large and complex (5 x human genome), and has been the target of a major international collaborative sequencing effort led by the International Wheat Genome Sequencing Consortium (IWGSC). The IWGSC included Australian research groups. A first draft of the complete 21 chromosomes of wheat was recently released to the wider community [1]. In this poster we describe the EMBL-ABR Activity: Wheat Annotation Platform, what technologies we have implemented, how it is working and its next steps.

The sequencing and assembly of chromosome 7A has been the Australian contribution to the IWGSC effort and a high-quality assembly is now reaching completion.

Defining regions of biological interest such as gene boundaries, splice variants and other important functional genomic features is a challenge and a number of automated approaches exist. However, in order to produce a truly high quality annotation, manual curation is required, as has been found to be the case in other genome projects. The EMBL-ABR infrastructure is establishing a generic annotation platform utilizing the wheat genome as a pilot project.

[1] http://www.wheatgenome.org/News/Press-releases/Wheat-Sequencing-Consortium-Releases-Key-Resource-to-the-Scientific-Community

Poster ID: 54 (AB3ACBS)

Bioinformatic investigation into effects of Wolbachia pipientis on insect-specific Flaviviruses populations in mosquitoes

Kirill Tsyganov (a), Hilaria Amuzu (b), Beth McGraw (b)

(a) Monash Bioinformatics Platform, Monash University
(b) School of Biological Sciences, Faculty of Science, Monash University

Insect-specific Flaviviruses (ISFs) belong to a group of insect-endogenous viruses that do not have the ability to infect vertebrates. Previously little attention has been paid to them but they are in fact important in the host's interaction with exogenous (dual-host) Flaviviruses which can cause disease in humans. Another interesting aspect of ISFs is their relationship with Wolbachia pipientis, an endosymbiotic bacterium of insects. This naturally occurring bacteria is found in 40% of arthropods, and is known to have suppressive effects on its host's reproduction and immunity, further it is a possible competitor for the same resources as ISFs. Again, there is currently little information about the positive or negative effect of Wolbachia on ISFs. It has been shown previously that in Drosophila melanogaster Wolbachia decreases the diversity of Flaviviruses in the host. However, our data appears to support the opposite effect. We have used deep amplicon sequencing from mosquitoes captured at three different environmental sites, two of which were previously targeted by Wolbachia release in the environment. We used the bioinformatics tool,vsearch, for classification of sequences into Operational Taxonomic Units (OTUs), of which we have found 83 sequence clusters. We then produced a number of visualisations of the data, some of the methods used include Shannon's entropy and Simpson's diversity index (SDI) as a measure of species diversity. We have also attempted to use statistical tests on SDIs to refute our null hypothesis that Wolbachia does not play a role in species diversity. With out current tests no statistically significant differences were found, however both plotting Shannon's entropy and SDI indicates an increase in species diversity. There is a potential confounding issue with differences between library sizes in the different environmental groups and literature does suggest that both methods can be susceptible to such libraries biases.

Poster ID: 55 (AB3ACBS)

Global Mapping of Bioinformatics Training in Australia

Sonika Tyagi (a), Dieter Bulach (b), Simon Gladman (b), Pedro Fernandes (c), Allegra Via (d), Jason Williams (e), Judit Kumuthini (f), Javier De Las Rivas (g,h), Andrew Lonie (b,i) and Maria Victoria Schneider (b,i)

(a) EMBL-ABR: AGRF Node, Australian Genome Research Facility, Melbourne, Australia
(b) EMBL-ABR: VLSCI Node, Victoria Life Science computation, Melbourne, Australia
(c) ELIXIR-PT, Institute Gulbenkian de Ciencia, Oeiras, Portugal
(d) ELIXIR-ITA, Italian Institute of Bioinformatics, Bari, Italy
(e) CyVerse, DNA Learning Center, Cold Spring Harbor Laboratory, New York, USA
(f) Centre for Proteomic and Genomic Research, Cape Town, South Africa
(g) Cancer research Centre, Salamanca, Spain
(h) Iberoamerican Society for Bioinformatics, Latin America
(g) EMBL-ABR HUB, University of Melbourne, Melbourne, Australia

Several efforts in postgraduate skilling up across a variety of bioinformatics skills for life scientists have emerged and continue to increase across the globe. Australia has also seen a steady increase in postgraduate bioinformatics training. Here we provide: (i) an overview on Bioinformatics training across Australia, (ii) a review on how this compares to what we see worldwide, (iii) a look on data we collected internationally on what is the perceived “most needed bioinformatics training” and how this is matched by the perception within Australia.

We then discuss how the particular special-temporal scale in Australia may influence the ability and accessibility of those based in Australia to efforts in bioinformatics training across the globe with respect to both trainers and end users.

We discuss this overview on terms of how to scale up, make the most of existing efforts within and outside Australia and in particular how EMBL Australia Bioinformatics Resource plans to work in this area in the coming year.

Poster ID: 56 (AB3ACBS)

PTMOracle: a Cytoscape app for co-visualising and co-analysing post-translational modifications in protein interaction networks

Aidan P. Tay, Chi Nam Ignatius Pang, Daniel L. Winter, Marc R. Wilkins

Systems Biology Initiative, The University of New South Wales, Sydney, New South Wales 2052, Australia
School of Biotechnology and Biomolecular Sciences, The University of New South Wales, Sydney, New South Wales 2052, Australia

The majority of eukaryotic proteins are modified by at least one post-translational modification (PTM). PTMs can occur on multiple sites of the same protein. This could be through the same type of PTM, or different types of PTMs. PTMs act as key regulators of protein activity, such as protein-protein interactions (PPIs). However, performing systematic investigations linking the functional role of PTMs to PPIs remain challenging. To address this, we have developed PTMOracle, a new Cytoscape app that facilitates the co-analysis and co-visualisation of PTMs in the context of PPI networks. With PTMOracle, users can develop systematic searches to identify proteins of interest, explore network visualisations to discover complex PTM-associated relationships and generate hypotheses about the role of PTMs in PPIs. In addition, PTMOracle also allows users to integrate other types of protein data such as the protein sequence and sequence annotations, including but not limited to domains, motifs and disordered regions, into PPI networks. To our knowledge, PTMOracle is one of the first tools to do this, allowing users to explore relationships between PTMs, sequence annotations and PPIs to better understand how PTMs might regulate PPIs. We illustrate how PTMOracle can be used to investigate PTM-associated relationships and their role in PPIs with case studies from yeast and human. PTMOracle is open-source and available on the Cytoscape app store: http://apps.cytoscape.org/apps/ptmoracle.

Poster ID: 57 (AB3ACBS)

Enrich2: a statistical framework for analyzing deep mutational scanning data

Alan F. Rubin (a,b,c,d), Nathan Lucas (e), Sandra M. Bajjalieh (e), Anthony T. Papenfuss (a,b,c,f,g), Terence P. Speed (a,g), Douglas M. Fowler (d,h)

(a) Bioinformatics Division, The Walter and Eliza Hall Institute of Medical Research, Parkville, VIC 3052, Australia.
(b) Department of Medical Biology, University of Melbourne, Melbourne, VIC 3010, Australia.
(c) Bioinformatics and Cancer Genomics Laboratory, Peter MacCallum Cancer Centre, Melbourne, VIC 3000, Australia.
(d) Department of Genome Sciences, University of Washington, Seattle, WA 98195 USA.
(e) Department of Pathology, University of Washington, Seattle, WA 98195 USA.
(f) Sir Peter MacCallum Department of Oncology, University of Melbourne, Melbourne, VIC 3010, Australia.
(g) Department of Mathematics and Statistics, University of Melbourne, Melbourne, VIC 3010, Australia.
(h) Department of Bioengineering, University of Washington, Seattle, WA 98195 USA.

Measuring the functional consequences of protein variants can reveal how a protein works or help unlock the meaning of an individual’s genome. Deep mutational scanning is a widely used method for multiplex measurement of the functional consequences of protein variants. A major limitation of this method has been the lack of a common analysis framework. We developed a statistical model for estimating variant scores that can be applied to many experimental designs. Our method generates an error estimate for each score that captures both sampling error and consistency between replicates. We apply our model to one novel and five published datasets comprising 243,732 variants and demonstrate its superiority, particularly for removing noisy variants, detecting variants of small effect, and conducting hypothesis testing. We implemented our model in easy-to-use software, Enrich2, that can empower researchers analyzing deep mutational scanning data.

Poster ID: 58 (AB3ACBS)

Identifying Human Papillomavirus in Head and Neck Squamous Cell Carcinomas

Alexandra L. Garnham (a, b), Kendrick Koo (a), Oliver Sieber (a, b), Gordon K. Smyth (a, c)

(a) The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, Victoria, Australia
(b) Department of Medical Biology, The University of Melbourne, Parkville, Victoria, Australia
(c) Department of Mathematics and Statistics, The University of Melbourne, Parkville, Victoria, Australia

Human Papillomavirus (HPV) is comprised of a large group of related viruses, with over 100 subtypes identified. It has been well established that many of the HPV subtypes cause cancers of the genitals and anus, as well as in some regions of the head and neck. HPV is detected in approximately 25% of all Head and Neck Squamous Cell Carcinomas (HNSC) with the majority of cases occurring in the oropharynx. While there is strong evidence connecting HPV as a major cause of oropharyngeal cancers, it has remained unclear if this same link exists between HPV and cancers of the oral cavity. To explore this possibility, we acquired raw RNA sequencing data from HNSC samples generated by The Cancer Genome Atlas. With these data, we have developed a pipeline utilizing the Subjunc aligner that enables us to deconvolve viral RNA from human at the subtype level. This allows us to detect the presence of viruses in the tumour samples. Using this pipeline we are now investigating genetic differences between HPV positive and negative samples in both oropharyngeal and oral cavity cancers. Furthermore, we are studying which viral genes are active in positive samples.

Poster ID: 59 (AB3ACBS)

Culture independent genome sequencing provides new insight into the microbiome associated with different types of diabetic foot ulcers (DFUs)

Sumeet Sandhu (a,b), Irani Udeshika Rathnayake (a,b) and Flavia Huygens (a,b)

(a) Institute of Health & Biomedical Innovation, Queensland University of Technology, Brisbane, Queensland, Australia
(b) Wound Management Innovation Cooperative Research Centre, Brisbane, Australia

Development of foot ulcers is a serious complication associated with diabetic patients worldwide. Approximately 25% of patients with diabetes will develop diabetic foot ulcers (DFUs) during their lifetime. Depending on the pathophysiology of DFUs, they are divided into different ulcer types such as neuropathic, ischaemic, neuro ischaemic and post-surgical. It is hypothesized that the microbial abundance and diversity varies greatly in each ulcer type and is contributing to delayed wound healing in diabetic subjects. This study examines the microbiome associated with different types of DFUs. Thirty diabetic patients with different DFU types were sampled over 24 weeks. Genomic DNA (gDNA) was extracted from patient samples and amplicon libraries were constructed targeting the V1 and V2 region of the prokaryotic 16S rRNA gene. The Ion Torrent Personal Genome Machine (PGM) was used for amplicon sequencing and the bioinformatics pipeline "mothur" was used to perform the quality control and analysis of the sequences obtained. Sequences containing ambiguous bases, more than six homopolymers, Phred score <25, primers, adaptors and barcode mismatches were removed. Chimera detection and removal was performed using UCHIME. The Calypso software program was used for further data mining, providing information on bacterial diversity and abundance in each DFU sample. The genome sequence analysis provided an insight to the significantly ($p<0.05$) differentially abundant bacterial genera present within different DFU types. This suggested that dominance of specific types of microbes in different DFU types can potentially be used as a biomarker to inform DFU outcome.

Poster ID: 60 (AB3ACBS)

From genes to pathways: differential expression and pathway analysis of RNA-Seq experiments

Yunshun Chen (a,b) and Gordon Smyth (a,c)

(a) The Walter and Eliza Hall Institute of Medical Research
(b) Department of Medical Biology, The University of Melbourne
(c) Department of Mathematics and Statistics, The University of Melbourne

In recent years, RNA sequencing (RNA-seq) has become a very widely used technology for profiling gene expression. One of the most common aims of RNA-seq profiling is to identify genes or molecular pathways that are differentially expressed (DE) between two or more biological conditions.

Here I will demonstrate a computational workflow for the detection of DE genes and pathways from RNA-seq data by providing a complete analysis of an RNA-seq experiment profiling epithelial cell subsets in the mouse mammary gland. It covers all steps of the analysis pipeline, including data exploration, differential expression analysis, visualization and pathway analysis. The analyses are performed using the edgeR package.

Poster ID: 61 (AB3ACBS)

KrakenDB* : A hybrid data repository for multi-omic querying of cancer data.

Wishva Herath, Ron Firestein

Centre for Cancer Research, Hudson Institute of Biomedical Research

Wide adoption of high-throughput methods such as RNA-Seq , have resulted in the rapid increase of cancer omics datasets. Efficient management and interrogation of these datasets is vital for the development of the field. Integration analysis studies which combine data from multiple omics experiments are of particular interest as they allow researchers to look beyond the confinements of a particular method and generate a comprehensive view of the biology of the sample-set. As of now, most repositories compartmentalize data by platform / technology which makes integrating omics data a time-consuming endeavour that also requires considerable bioinformatic talent.

To solve this problem, we designed KrakenDB a repository built ground up for performing multi-omic queries. KrakenDB has a hybrid data structure made up of tabular data which is linked by a graph/network - created using existing knowledge and available metadata. The hybrid data structure allows more flexibility when expressing relationships within the database compared with the offerings of traditional relational databases. A query in KrakenDB first involves a network traversal followed by extracting values from tables. A query can either be issued by directly issuing traversal commands or by a simple ‘english-like’ query language.

KrakenDB was designed to be “biology aware” and has built-in information such as gene synonyms, gene co-ordinates etc which makes queries simpler. KrakenDB is written in python.

Poster ID: 62 (AB3ACBS)

Rna-Seq: The effect of Single-end vs Paired-end and Stranded vs Nonstranded sequencing

Susan Corley (a) Marc Wilkins (a)

(a) Systems Biology Initiative, School of Biotechnology and Biomolecular Sciences, UNSW

RNA-Seq technology is capable of detecting all forms of RNA transcribed from the genome (including mRNA coding for proteins, microRNA, snoRNA, lincRNA, and mtRNA) and is now widely used by research groups as a valuable research tool. A number of decisions need to made prior to sequencing, including the appropriate number of replicates to sequence, blocking strategies, how to filter the RNA prior to sequencing and then whether to use a stranded library preparation kit and whether to employ single-end or paired-end sequencing. The decisions taken at this point will affect the cost of the RNA-Seq experiment and the utility of the data produced. Here we compare the outcomes of read mapping, feature counting and differential expression analysis using different experimental protocols involving single-end or paired-end sequencing and a stranded or non-stranded library preparation protocol. We explore these issues in four experiments involving human or mouse samples. Library preparation and sequencing was performed at the Ramaciotti Centre for Genomics. Reads were mapped to the human or mouse reference (Ensembl GRCh38 or Ensemble GRCm38) using Tophat2and then to features using Subread. Differential expression analysis was performed using edgeR and limma (voom).

Poster ID: 63 (AB3ACBS)

High-Resolution Termite Metagenomics

Boyd Tarlinton, Christopher Noune, Caroline Hauxwell

Queensland University of Technology

The use of meta-barcoding enables high-resolution insights into the metagenomes of various organisms that cannot be accomplished with conventional wet-lab techniques or shot-gun ‘next generation sequencing’ (NGS). Targeted NGS of the taxonomically significant 16S rRNA meta-barcode was used to identify and determine relative abundance of bacterial strains within the termite microbiome. A recently developed meta-barcoding software pipeline called ‘IMG-GAP’ was used to produce a high-resolution map of these bacterial strains. This pipeline produces a sequence database based on polymorphisms identified within a dataset, enabling the production of meta-barcodes that are more highly resolved than those produced with pipelines that cluster reads based on sequence similarity. It was revealed by this analysis that the samples were dominated by bacteria of the phylum Fibrobacteres. Also revealed by the analysis was the greater degree of bacterial diversity that exists within termites of the worker caste compared to soldier termites of the same species.

Poster ID: 64 (AB3ACBS)

SMALL RNA PROFILING USING DIFFERENT SAMPLE PREPARATION PROTOCOLS AND INPUT AMOUNTS

Sonika Tyagi, Jafar Jabbari, Lavinia Gordon, Matthew Tinning, Kirby Siemering

Australian Genome Research Facility Ltd.

Recent advances in high throughout sequencing have aided the discovery of new classes of small non-coding RNAs and understanding their roles in gene regulation. Current small RNA sample preparation protocols predominantly target the RNAs in the range of 18-25nt length, of which microRNAs are the most abundant (~40-50 %) RNA species. Different sample preparation protocols come with their possible bias between relative expression levels and sequences. Care must be taken when choosing a particular protocol, especially for experiments with limited input RNAs, for more reliable and accurate smallRNA detection. A given protocol may not fit all experimental conditions and available inputs and currently there are no gold standards for small RNA library preparation. In this paper we provide a comparative analysis of two leading library preparation protocols; TrueSeq from Illumina and NEBNext Multiplex from New England BioLabs (NEB), respectively at various input RNA levels (1000ng, 500ng and 100ng). We compared how two protocols influenced sequencing yields, quantitative measurement of miRNA abundance, detection of lowly expressed RNAs and reproducibility at different inputs. We compared differential expression, reproduciblity and correlations between various inputs. We did not see one preparation absolutely outperforming the other however, the results indicate strengths and weakness of the two protocols and variability in outcome introduced by different amount of starting material.

We believe that such bench marking studies will help researchers make an informed decision about the sample preparations that best fit their study objective.

Poster ID: 65 (AB3ACBS)

Comprehensive evaluation of the molecular and cellular activity of therapeutic small molecule BET inhibitors

Enid Y. N. Lam (a,b), Dean Tyler (a,b), Johanna Vappiani (c), Tatiana Cañeque (d), Omer Gilan (a,b), Aoife Ward (c), Yih-Chih Chan (a), Antje Hienzsch (d), Chun Yew Fong (a,b), Laura MacPherson (a), Sarah-Jane Dawson (a,e), Gerard Drewes (c), Rab K. Prinj

(a) Cancer Research Division, Peter MacCallum Cancer Centre, Melbourne, Australia
(b) The Sir Peter MacCallum Department of Oncology, University of Melbourne, Melbourne, Australia
(c) Cellzome GmbH, Molecular Discovery Research, GlaxoSmithKline, Heidelberg, Germany
(d) Institut Curie, PSL Research University, Paris, France (e) Centre for Cancer Research, University of Melbourne, Melbourne, Australia
(f) Epinova DPU, Immuno-Inflammation Therapy Area Unit, GlaxoSmithKline, Stevenage, United Kingdom
(g) Department of Haematology, Peter MacCallum Cancer Centre, Melbourne, Australia

The development of targeted therapies to specifically eradicate malignant cells has transformed the treatment of a number of cancers. However, most of the novel therapies that have shown early promise in pre-clinical research fail to progress to the clinic. Underpinning this failure to succeed in the clinical domain is a lack of knowledge of the molecular and cellular effects of the novel therapy.

BET bromodomain inhibitors (BETi) are a class of novel cancer therapeutics that target the bromodomains of the epigenetic BET proteins leading to their displacement from chromatin, and are currently being investigated in clinical trials for haematologic and other cancers. By modifying various BET inhibitors, we created functionally preserved compounds that are amenable to click-chemistry and can be used as molecular probes in vitro and in vivo. Using click-sequencing, we identified regions of the genome that are targeted by BETi.

At transcription start sites, where BRD4 plays an important role in transcription regulation, only a minority of genes are responsive to BETi. We found that while BRD4 is ubiquitous at the TSS of transcribed genes, the levels of BETi at the TSS correlate with the responsiveness of a particular gene to transcriptional inhibition. At enhancer regions, there is also differential chromatin drug occupancy relative to BRD4. Motif analysis showed an enrichment of known BRD4–interacting transcription factor binding sites at these drug inaccessible enhancers. The differential accessibility of BRD4 to BETi in different genomic regions suggest multiple mechanisms of BRD4 binding to chromatin that provide new insights to explain the gene regulatory function of BRD4 and the transcriptional changes invoked by BET inhibitors.

Poster ID: 66 (AB3ACBS)

Integrative genomics of circulating metabolites in human populations

Aaron Casey (a), Song Gao (a), Ville-Petteri Mäkinen (a, b, c)

(a) Heart Health, South Australian Health & Medical Research Institute, Adelaide, South Australia
(b) School of Biological Sciences, University of Adelaide, Adelaide, South Australia
(c) Computational Medicine, Faculty of Medicine, University of Oulu and Biocenter Oulu, Oulu, Finland

Molecules in blood are important indicators of metabolic health, but the genetics of circulating metabolites and their relevance for human diseases are not fully understood. To elucidate the genetic regulation of metabolism, we conducted a pathway enrichment and network analyses of the GWAS summary statistics for 123 metabolic traits in a Northern European consortium of 24,925 individuals. Curated pathways were extracted from MSigDB v5.0 (Broad Institute, USA), drug signatures from DSigDB (University of Colorado, USA) and disease signatures from GWAS Catalog (European Bioinformatics Institute, UK). Protein-protein interaction network was obtained from STRING v10 (STRING Consortium, Europe). The data were analysed by Mergeomics (University of California Los Angeles, USA). The lipoprotein metabolism pathway (P = 2.06x10-11 for top signal), cholesterol-lowering drugs (P = 5.99x10-09) and cardiovascular disease (P = 5.23x10-12) were genetically associated with circulating lipoprotein subclasses, as expected. Genetic perturbations to circulating glycine were enriched in: ERBB signalling pathways (P = 5.93x10-06 for top signal); pd 158780 drug signature (P = 1.81x10-05); Tyrphostin signature (P = 2.45x10-05), and; breast cancer (P = 5.98x10-04). Network analyses indicated apolipoproteins A1, E and B as top key drivers of lipoprotein metabolism (P < 2.08x10-23). ERBB2 was the top key driver for the ERBB signalling pathways (P = 2.06x10-11) (ERBB3 amongst the top key drivers P = 5.30x10-11). CCND1 (P = 6.25x10-13) and ERBB3 (P = 5.99x10-09) were top key drivers for the pd 158780 drug signature, with MAX (P = 1.07x10-06) the top key driver for breast cancer (ERBB3 also a top key driver P = 1.32x10-02). The integrative analysis of metabolome GWAS identified both expected and unexpected patterns of genetic associations. Circulating glycine may indicate increased genetic risk for breast cancer, although further epidemiological studies are needed to confirm the association.

Poster ID: 67 (AB3ACBS)

Biomarker selection incorporating data independent acquisition mass spectrometry for the prediction of response to treatment in a non-healing wound cohort

Daniel A. Broszczak (a, b), James A. Broadbent (a), Dayle L. Sampson (a), Zee Upton (a, c), Tony J. Parker (a, b)

(a)Tissue Repair & Regeneration Program, Institute of Health and Biomedical Innovation, Queensland University of Technology, Kelvin Grove, Australia
(b)School of Biomedical Sciences, Faculty of Health, Queensland University of Technology, Brisbane, Australia
(c)Institute of Medical Biology, Agency for Science, Technology and Research (A*STAR), Singapore

Chronic wounds are highly complex open sores that take months or years to heal, the reasons for which remain largely unknown. Understanding the underlying biochemistry of chronic ulcers may provide much needed insight into why these wounds are recalcitrant. Moreover, proteins present within wound fluid may be indicative of whether an ulcer will heal given best-practice clinical care. Using multiple fractionation methods to separate proteins by various physiochemical properties prior to liquid chromatography tandem mass spectrometry we were able to establish a comprehensive wound fluid proteome spectral library. Multiple methods have provided the means to substantially increase the number of proteins detected within samples. Gene ontology enrichment of the proteome revealed an underlying biochemical difference between chronic wounds that heal compared to those that do not given best-practice treatment. Subsequently, we have performed data independent acquisition mass spectrometry (SWATH-MS) to quantify abundance changes in the proteome of wound fluids collected from chronic ulcers. These data were analysed using permutation statistics and a novel regression algorithm that determined a suite of biomarkers. These markers can be used to classify ulcers based on their expected healing outcome within 24 weeks of best-practice clinical care. Collectively, these data demonstrate the value of studying wound fluid to better understand ulcers and, furthermore, underscore the viability of wound fluid as a non-invasive medium for monitoring wounds in clinical practice.

Poster ID: 68 (AB3ACBS)

BayesMS: A Bayesian Model Selection approach for identifying cell- and sex-dependent DNA methylation patterns, accounting for cell lineage in whole blood.

Nicole White (a), Miles Benton (b), Daniel Kennedy (a), Andrew Fox (c), Lyn Griffiths (b), Rod Lea (b), Kerrie Mengersen (a)

(a) ARC Centre of Excellence for Mathematical and Statistical Frontiers, Queensland University of Technology (QUT), Brisbane, Queensland, Australia
(b) Genomics Research Centre, Institute of Health and Biomedical Innovation, Queensland University of Technology, Brisbane, Queensland, Australia
(c) The Florey Institute of Neuroscience and Mental Health, Parkville, Victoria, Australia

DNA methylation is an epigenetic mechanism associated with the regulation of gene expression. The analysis of differential DNA methylation with respect to cell type and its relationship with biological pathways has received significant attention. However, strategies for the identification of cell-dependent patterns at the individual CpG level ("CpG markers") have to date been restricted to a single cell type and have ignored potential lineage effects. Furthermore, sex-specific analysis of markers identified by these methods remains relatively unexplored.

This work proposes a Bayesian model selection algorithm (BayesMS) for the discovery of single and multi- cell-dependent methylation profiles, with the goal of establishing CpG marker panels for major leucocytes subtypes (CD19+ B, CD4+ T, CD8+ T, CD14+ Monocytes, CD56+ NK, CD16+ Neutrophils). The approach involves identification of cell-dependent CpG markers by using prior knowledge of shared hematopoietic cell lineages. The model selection process accounts for uncertainty in both cell-dependent methylation levels and the true methylation pattern at each CpG.

The BayesMS algorithm was applied to cell-sorted Illumina 450K methylation data on 11 subjects (6 male, 5 female) allowing for marker panels for each sex, and their intersection, to be derived. In total, 42,452 cell-dependent CpG markers were identified as common between female and male samples. Of these markers, 23,551 (55.5%) were associated with differential methylation in more than one cell type. Variation in the magnitude of differential methylation increased with the complexity of the cell-dependent pattern. Sex-specific differences in methylation among common markers were generally low, however, higher methylation in males between 13.8 and 29.1% in CD16+ Neutrophils and between 13.3 and 35.1% (95%CI) in CD56NK+ was observed in select marker panels. Enrichment analysis of identified markers revealed significant associations with relevant signaling pathways.

Poster ID: 69 (AB3ACBS)

Flexible analytical pipelines for large virus datasets

Jan P Buchmann, Mang Shi and Edward C. Holmes

School of Life & Environmental Sciences, The University of Sydney, Sydney, AUS

Virology is entering a ‘discovery phrase’. Because of new and cheaper sequencing technologies, particularly RNA-Seq, a deluge of transcriptome data is becoming available from a wide range of organisms. Crucially, not only do these sequence data provide information on the (eukaryotic) organisms of interest, but they also contain a record of all the expressed microbial parasites that infect that host, including RNA viruses. Despite the power of these data, they are usually only analyzed in the context of the given project, and rarely considered in the context of viruses and their evolution. That this constitutes a major missed opportunity is evident from the fact that most studies of virus biodiversity and evolution have focused on either specific types of virus or specific hosts, such that our knowledge of virus diversity in most eukaryotic taxa is scant. To be able to analyze such rich datasets, we are working on several small tools which can integrate the results of existing analysis tools to perform specific evolutionary analysis. This distributed approach increases the flexibility and adaptability of analysis pipelines while avoiding the complexity of a monolithic approach.

Poster ID: 70 (AB3ACBS)

Survival Volume : A Python package for interactive tumour volume plots

Matthew J. Wakefield (a,b)

(a) Bioinformatics, The Walter and Eliza Hall Institute
(b) VLSCI, The University of Melbourne

Volume measurement over time is a common measure of treatment effectiveness in preclinical and clinical research. This data is often very rich, and can contain valuable information in addition to the usual time to endpoint that contributes to Kaplan-Meier curves. Common practice is to present this data as three unrelated plots: a 'spaghetti plot' of all individuals, a mean plot with error, and a Kaplan-Meier curve.

Survival Volume is a high level python plotting package that allows creation of aesthetically pleasing plots that combine all these features and allow the data to be clearly explored. Alpha transparency scaling is used to display the mean, confidence interval, and individual subjects on the same axis while preserving legibility. A Kaplan-Meier plot stacked vertically and sharing the x axis scale provides an alternative representation of the data while maintaining association between the events depicted in both panels. Additional clarity of the trajectory of individual subjects can be obtained using mouseover support, highlighting the line plot for an individual and providing a sample label.

Survival Volume utilises the matplotlib, mpld3 and lifelines python packages and is designed for stand alone or jupyter notebook use. The software is licensed under the GPLv3 and available from http://github.com/genomematt/SurvivalVolume

Poster ID: 71 (AB3ACBS)

Accelerating alignment: what can we gain from GPU-based sequence alignment tools?

Sean Li (a), Ondrej Hlinka (b), Annette McGrath (a)

(a) CSIRO Data61, GPO Box 664, Canberra, ACT 2601
(b) CSIRO IMT, QBP, St Lucia, QLD 4072

The development of efficient short-read alignment tools has still been considered as a challenging topic in Bioinformatics due to the largely accumulated short sequence reads generated from the very high-throughput but low-cost next generation sequencing (NGS) platforms. In the past few years, a number of graphic processing units (GPU)-based short-read aligners have been proposed to fully utilize the parallel power of GPUs (e.g., SOAP3-dp, cushaw2-GPU and Barracuda), which aim to further accelerate the alignment process without the loss of accuracy in contrast to the CPU ones. Additionally, NVidia has implemented bowtie (NVbowtie) for GPUs. This GPU implementation therefore affords the opportunity to make a direct comparison of tool's performance under CPU vs GPU. Here, we have carried out a comparative study to investigate the capabilities of GPUs and GPU-aware short-read aligners against CPU ones to understand in what circumstances it would be beneficial to use GPUs and what limitations there are in our current GPU infrastructure. The main facets to examine are speed, accuracy and data size. Moreover, we compared the performance of nvBowtie over multiple GPUs on both Tesla K20s and K80s with larger datasets. As alignment is often the first step in downstream analysis, using the Genome In A Bottle datasets we have evaluated downstream variant detection using GATK for both CPU and GPU-based aligners. This poster outlines the datasets and tools we chose, the experimental workflows, preliminary outcomes and related experience we have learnt.

Poster ID: 72 (AB3ACBS)

Using a weighted bootrap to detect the role of transposable elements on global gene regulation

Stephen Pederson, Lu Zeng, David Adelson

University of Adelaide

Nearly half of the human genome is made up of transposable elements (TEs) and evidence supports a possible role for TEs in gene regulation. Here, we have integrated publicly available genomic, epigenetic and transcriptomic data to investigate this potential function in a genome.-wide manner.

Through the use of a weighted bootstrap, our results show that while most TE classes are primarily involved in reduced gene expression, Alu elements are associated with higher levels of gene expression. Furthermore, non-coding regions were found to have a greater density of TEs within regulatory sequences, most notably in repressors. This exhaustive analysis has extended and updated our understanding of TEs in terms of their global impact on gene regulation, and indicates a significant association between repetitive elements and gene regulation.

Poster ID: 73 (AB3ACBS)

The Genomic Landscape of Oesophageal Adenocarcinoma and its precursor Barrett's Oesophagus

Felicity Newell (a), Katia Nones (a), Kalpana Patel(d), Michael Gartside(d), Lutz Krause(b), Kelly A. Loffler (d), Stephen Kazakoff (a), Ann Marie Patch (a), Luke F. Hourigan (d), Bradley J. Kendall (a,e,f), David C. Whiteman (a), John V. Pearson (a), Nicola Waddell(a), Andrew P. Barbour (c,d)**

(a) QIMR Berghofer Medical Research Institute, 300 Herston Road, Herston 4006, Brisbane, QLD Australia
(b) The University of Queensland, Diamantina Institute, Translational Research Institute, Woolloongabba 4102, Brisbane, QLD, Australia
(c) Department of Surgery, School of Medicine, The University of Queensland, Princess Alexandra Hospital, Woolloongabba 4102, Brisbane, QLD, Australia
(d) Surgical Oncology Group, School of Medicine, The University of Queensland, Translational Research Institute at the Princess Alexandra Hospital, Woolloongabba 4102, Brisbane, QLD, Australia
(e) Department of Gastroenterology and Hepatology, Princess Alexandra Hospital, Woolloongabba 4102, Brisbane, QLD, Australia
(f) School of Medicine, The University of Queensland, Princess Alexandra Hospital, Woolloongabba 4102, Brisbane, QLD, Australia

Oesophageal adenocarcinoma (EAC) is associated with poor survival, with less than 20% of patients surviving for 5 years. The precursor to EAC is metaplastic Barrett's oesophagus, however the molecular characteristics that are associated with the progression from Barrett's to EAC are not yet fully understood. In this study we used whole genome sequencing to examine the genomic landscape of 57 EAC and 22 Barrett's oesophagus samples. Barrett's oesophagus samples included samples from patients who did not progress to EAC, as well as non-dysplastic and high grade dysplastic Barrett's samples. The mutational burden of somatic SNP and indel mutations was similar between high grade dysplastic Barrett's (median of 5.8 mutations per megabase) and EAC samples (median of 5.3 mutations per megabase) but lower in non-progressor and non-dysplastic Barrett's samples. We observed the presence of five mutational signatures in EAC, including the T>G at TT sites signature that is common in EAC and is thought to be a result of exposure to bile and gastric acids. This signature was also found to be present in Barrett's samples. Significantly mutated genes in EAC included TP53, CDKN2A and SMAD4, with mutations for TP53 and CDKN2A also present in Barrett's samples. Copy number and structural variations were frequently observed in EAC samples, including recurrent amplifications across the genome. Copy number variations were low in Barrett's samples, particularly in terms of amplification events. These observations agree with previous studies that some of the molecular events associated with EAC are already present in the Barrett's oesophagus precursor phase.

Poster ID: 74 (AB3ACBS)

Glimma, getting greater graphics for your genes

Shian Su (a), Charity W. Law (a,b) and Mathew E. Ritchie (a,b,c)

(a) The Walter and Eliza Hall Institute of Medical Research, Parkville, 3052, Australia
(b) Department of Medical Biology, The University of Melbourne, Parkville, 3010, Australia
(c) School of Mathematics and Statistics, The University of Melbourne, Parkville, 3010, Australia

RNA-sequencing is a popular technology used by scientists to study changes in gene expression levels across tens of thousands of genes simultaneously. Representing gene expression levels, the counts in each sample are typically analysed by categorising samples into groups of interest, and obtaining gene-wise summary statistics in the form of log-fold changes, t-statistics, p-values, and the like. The data and its results can be explored by plotting one summary statistic against another and highlighting genes that are significant or of interest. The new Bioconductor package, Glimma, generates interactive graphics for plots typically found in the limma package with the enhanced feature of connecting many levels of information within the analysis on a single html page using d3.js. A Glimma-style mean-difference plot, or the more generic xy-plot, allows one to click on the points to bring up a new plot of sample-wise expression levels that is displayed alongside the original plot. This feature enables researchers to interrogate the data more intensely than ever before without the need to repeat the work for every gene under examination. The plots include options to search and select for genes of interest, and zoom in and out for better resolution. Unlike the traditional multi-dimensional scaling (MDS) plot, Glimma's MDS plot shows several dimensions and group combinations on the same page. The functions within Glimma are tailored to integrate smoothly with objects native to limma, edgeR and DESeq2, and can be extended for use with microarray, single-cell and methylation data analyses.

Poster ID: 75 (AB3ACBS)

Scaling bioinformatics trainers: How to meet the demand for bioinformatics training

Annette McGrath (a), Katherine Champ (a), Konsta Duesing (a), Paul Greenfield (a), Sean McWilliam (a), Paula Moolhuizen (a), Erdahl Teber (a), Sonika Tyagi (a), Sarah Morgan (b)

(a) BPA/CSIRO Training Network
(b) EMBL-EBI, Hinxton, Cambridgeshire, UK

The widespread adoption of high-throughput next-generation sequencing (NGS) technology among the Australian life science research community is highlighting an urgent need to up-skill biologists in tools required for handling and analysing their NGS data. Bioinformaticians come from a wide range of academic backgrounds and are often required to gain skills in handling new types of ‘omics data. In addition, rapid ‘omics technology churn further adds to the demand for on-going training in new tools and analysis methods.

Commonly, bioinformatics trainers are volunteers who run training courses on a sporadic basis. As a consequence of a scarcity of skilled trainers with time and allocated funding to develop and deliver training courses, there are still gaps in bioinformatics training provision nationally.

Since 2012, groups of Australian bioinformaticians have attended train-the-trainer workshops at the European Bioinformatics Institute (EBI) and at The Genome Analysis Centre (TGAC) to improve their training skills in developing and delivering bioinformatics workshop curriculum. Some trainers have attended multiple train-the-trainer sessions and have subsequently delivered bioinformatics training courses nationally to over 1000 students. T

This experienced group are now undertaking a pioneering Train-the-trainer (TtT) Instructor course in conjunction with the EMBL-EBI node of ELIXIR, the Europe-wide bioinformatics infrastructure project, and in conjunction with GOBLET, The Global Organisation for Bioinformatics Learning, Education and Training, to further scale bioinformatics training activities in Australia. On completion of the TtT Instructor course, this group will deliver a Train-the-trainer course to a group of bioinformaticians at the AB3ACBS conference and workshop who wish to develop their training skills, thus increasing the pool of those skilled in delivering bioinformatics training nationally.

Poster ID: 76 (AB3ACBS)

Soft selective sweeps on polygenic traits in human populations

Emily Wong (a), Matthew Robinson (b), Joseph Powell (b)

(a) School of Biological Sciences, University of Queensland
(b) Institute for Molecular Biosciences, University of Queensland

Genomic regions that have undergone positive selection during species evolution are indicative of functional adaptations that drive population differences. We measured genetic differentiation at variants across global human populations to detect evidence of adaptive selection affecting GWAS traits and examined how natural selection has acted upon loci shared in multiple traits. Shared loci show elevated levels of natural selection. We find evidence for balancing selection acting at pleiotropic loci and show that selective forces at pleiotropic loci can promote positive selection and adaptation, rather than hindering it.

Poster ID: 77 (AB3ACBS)

Long k-mer clustering for scalable and accurate biological search

Timothy Chappell (a), Lawrence Buckingham (a), Shlomo Geva (a), Paul Greenfield (b), Wayne Kelly (a), Jim Hogan (a)

(a) Queensland University of Technology
(b) CSIRO

BLAST, the Basic Local Alignment Search Tool, remains the dominant method for general purpose search and sequence comparison in molecular biology. While a number of alternatives – including some based directly on BLAST itself – are significantly faster, this efficiency may come at the cost of sensitivity, with performance degrading sharply for more divergent targets. In many cases, the underlying representation - through k-mer indices and similar approaches – proves highly effective at finding close matches, but is unable to discriminate sufficiently to find matches in the so-called protein twilight zone.

In this work we introduce an approach based on hierarchical clustering of long protein k-mers into a k-mer database. Search is then realised by successively removing the most distant clusters from consideration and combining the similarity scores and position information from remaining k-mers. This allows us to produce a similarity score and ranking that contains all but the least likely matches. The resulting algorithm is several times faster than NCBI BLASTP with in many cases equal or greater precision.

Poster ID: 78 (AB3ACBS)

Cross-Cultivar Transcriptomics: A novel method for comparing pathogen-induced gene expression between non-model and polyploid host varieties.

Adam P. Taranto, Lauren Du Fall, Megan C. McDonald, Peter. S. Solomon

Plant Sciences Division, Research School of Biology, The Australian National University, Canberra, Australia

Following infection, gene expression of a plant host can reveal pathways and processes targeted by a successful fungal pathogen. Equally, plant genes deployed in a successful defence response may betray fungal weaknesses, and provide clues for breeding durable resistance. Accurate prediction of differentially expressed genes between cultivars relies on the identification of equivalent transcripts. For many crop species, reference genomes are not available and generation of near-isogenic lines is impractical; necessitating the use of de novo transcriptome assemblies. Traditional identity-based clustering methods face a trade-off in polyploid species; strict clustering allows true divergent homologues between cultivars to group together but risks over-clustering of biologically relevant homeologs, paralogues and splice variants.

We employ a recently developed hierarchical, shared-read, clustering method to assign equivalent transcripts between independently assembled de novo transcriptomes. This method is superior to pooled-read assemblies, and allows preservation of splice variants with distinct expression profiles while minimising homeologue collapse. Two allohexaploid wheat (Triticum aestivum) cultivars, with differential susceptibility to the necrotrophic fungal pathogen Parastagonospora nodorum were infected, and their transcriptomes sequenced. This work highlights several known components of anti-fungal plant defence at play in susceptible and non-susceptible varieties, as well as suggesting a role for novel candidates.

ADVANCE QUEENSLAND

made for innovation

Advance Queensland is a comprehensive suite of programs that will create jobs now and jobs for the future, drive productivity improvements, and harness innovation.

It will help Queensland attract and retain the best and brightest minds and nurture a culture that supports entrepreneurialism to flourish.

To capitalise on the opportunities of tomorrow, we need to:

- ensure our young people have the skills that will be in demand—science, technology, engineering and maths, including computer coding abilities
- support our graduates, new businesses and scientists to think and network globally
- encourage industry and research organisations to work together to translate more scientific and technological excellence into products and services
- work with the private sector to raise the rate of startups and better support businesses to access the advice, mentoring and capital to grow.

Advance Queensland will continue to build Queensland's reputation and capacity to conduct innovative research and development that translates into real outcomes.

We want Queensland recognised globally as a place where industry, researchers, and government work collaboratively to take great ideas through proof of concept to investment-ready products and businesses.

An expert panel has been established to support the implementation of Advance Queensland. The expert panel is chaired by the Minister for Innovation, Science and the Digital Economy and Minister for Small Business, the Hon. Leeanne Enoch MP, and the Deputy Chair is the Queensland Chief Scientist Dr Geoff Garrett. The expert panel brings together successful leaders from across the business, academic, research and education sectors to provide broad-based expertise and independent advice to government.

Advance Queensland includes:

- $50 million Advance Queensland Best and Brightest Fund: develops, attracts and retains world-class talent, both scientific and entrepreneurial.
- $46 million Advance Queensland Future Jobs Strategy: opens the door to new industry–research collaborations, tackles the big innovation challenges, focuses on translation, and delivers 10-year roadmaps for industries with global growth potential.
- $76 million Advance Queensland Business Investment Attraction package: encourages a new wave of Queensland startups, supports proof-of-concept projects and attracts co-investment through the Business Development Fund.

160219 AQ October 16

Queensland Government

Investing in our future

Developing, attracting and retaining world-class talent and skills

Queensland is globally recognised for its great talent, and we have invested heavily over many years in science and research skills. We want to continue to invest in this talent to keep the bright ideas, innovators and entrepreneurs in Queensland to benefit our economy.

The $50 million Advance Queensland Best and Brightest Fund will help develop, attract and retain world-class talent—both scientific and entrepreneurial.

- We will invest in a range of **fellowships and scholarships** at our universities and research institutions that specifically focus on researchers, women, Aboriginal and Torres Strait Islander students, and regional locations. These are:
  - Advance Queensland Research Fellowships to support postdoctoral research fellowships that focus on links with industry, supporting women, regions and researchers.
  - Advance Queensland PhD Scholarships to attract and retain promising researchers in Queensland by supporting undergraduates in gaining a research PhD degree. It will also encourage increased links and closer collaboration with industry/end-user organisations.
  - Advance Queensland Aboriginal and Torres Strait Islander Research Fellowships to address under-representation by offering funding to early-career researchers.
  - Advance Queensland Aboriginal and Torres Strait Islander PhD Scholarships to lay the fundamental foundation for a future research career.
- **Advance Queensland Masters Scholarships** will foster greater participation in key disciplines by female researchers, Aboriginal and Torres Strait Islander researchers, and researchers from low socioeconomic backgrounds.
- **Advance Queensland Global Partnership Awards** will support collaboration between Queensland and international researchers and entrepreneurs. This will offer graduates, researchers and emerging entrepreneurs the chance to learn directly from overseas successes.
- **Advance Queensland Knowledge Transfer Partnerships** program will link industry and universities and promote the use of research and problem-solving skills to assist business growth. Small to medium enterprises will have graduate students work in their company on strategic projects to help develop their products or services.
- We will undertake a **future schools review** of the teaching of STEM (science, technology, engineering and maths) in Queensland schools. This will include how to expand the program to include coding and computer science, as well as early-stage robotics and entrepreneurial skills. We will also transform the teaching of STEM through focused professional development, teacher scholarships and by working more closely with universities.
- **Advance Queensland Women's Academic Fund** supports women in maintaining their research careers, and supports the promotion of the achievements of their female researchers. It aims to retain, develop and progress female researchers within Queensland-based universities and publicly funded research institutes or organisations.

Translating ideas and research

Priority industries and technologies with global growth potential

The $46 million Advance Queensland Future Jobs Strategy will build on the Queensland Government's previous investment in research, and focus on building a culture of collaboration between research bodies and business to translate ideas and research into products, processes, service outcomes and jobs. Major investments in research infrastructure by previous governments through the Smart State initiative have laid a great foundation for the Advance Queensland focus on people and projects, and effective translation to commercial outcomes, new and growing businesses, and jobs.

Flagship partnerships

- Queensland Government has secured three major **flagship partnerships** that will help put the spotlight on Queensland and deliver the ultimate goal of turning ideas into jobs:
  - The Queensland Emory Drug Discovery Initiative, a collaborative project between The University of Queensland and Emory University in Atlanta, USA, will take Queensland researchers' ideas for potential cures for diseases and develop these into new drugs. Emory has an impressive record in drug discovery and is responsible for the development of the world's leading HIV treatments.
  - The Siemens Innovation and Translation Centre at the Translational Research Institute is a world leader in the field of MRI scanning technology. This innovation centre will ensure that Queensland has access to state-of-the-art scanning technology, and will train highly skilled technicians to operate and program these tools.
  - The Johnson & Johnson Innovation Partnering Office@QUT will facilitate access to the vast resources and expertise across Johnson & Johnson's scientific research, investor and commercial business sectors to help build, nurture and accelerate the local life science ecosystem.
- As well as the discoveries and advances these projects will create, they will be a magnet for innovators around the globe to consider Queensland as a potential location for business.
- **Advance Queensland Innovation Partnerships** will support collaborative research and development projects involving both research organisations and industry, to address industry issues in priority areas such as agriculture, engineering, climate change, clean energy, biotechnology and advanced manufacturing.
- **Advance Queensland Innovation Challenges** will harness the attention of researchers and industry to work collaboratively with government to address big issues and opportunities facing Queensland. The process will be seeking to make breakthroughs that may not otherwise be achieved. This is a globally successful model where our best innovators can rise to the challenge.
- **Advance Queensland Supporting Priority Industries Program** will invest in 10-year roadmaps where government will work with industry, academic and research partners to develop long-term plans for emerging industries, including medical technical, industrial biotechnology and biofuels.

The Advance Queensland Future Jobs Strategy places Queensland at the Australian forefront for funding industry–research collaborative projects. In doing so it helps build many dynamic and successful industry-savvy researchers, making Queensland the destination of choice for businesses looking to partner with an excellent research base.

Boosting our entrepreneurial culture

Supporting Queensland startups and the growth of small to medium enterprises

Key drivers of economic change and jobs growth are entrepreneurs and ambitious businesses.

The $76 million Advance Queensland Business Investment Attraction package will improve access to finance and management support for startups and small to medium enterprises.

- The $24 million **Startup Queensland Program** will increase the rate of startup formation, attract more businesses to Queensland and encourage growth in existing small to medium enterprises. Startups can reshape entire industries through technology and business model innovation. They are vital to job creation and prosperity and help ambitious businesses to access the support needed for accelerated growth.

  We want to work with existing incubators and accelerator programs to provide an integrated suite of seed funding, co-working space, mentoring, and connection to customers and markets.

  This will also include a pilot **Queensland Small Business Innovation Research** initiative based on highly successful models in the USA and UK.

  This competition-based program will generate new business opportunities for small to medium enterprises, a route to market for their ideas, and bridge the seed funding gap experienced by many early stage companies. Additionally this will help bring more innovation into government procurement.
- The $12 million **Advance Queensland Commercialisation Program** will support proof-of-concept projects, designed to lead to new products and services.
- The $40 million **Business Development Fund** will provide co-investment to match and encourage greater angel and venture capital investment in Queensland businesses. It will help many Queensland businesses translate their ideas into products and services and access global export markets.

Further information

🌐 **advance.qld.gov.au**
✉ advancequeensland@dsiti.qld.gov.au
☎ 13 QGOV (13 74 68)

Join the conversation
#AdvanceQld

ACKNOWLEDGEMENTS

Queensland University of Technology would like to thank the following organisations who have generously supported the COMBINE16 Symposium and AB^{3}ACBS 2016 Conference:

Gold partners

illumina®

ADVANCE QUEENSLAND

Silver partners

SARSTEDT

westmead research hub

QIMR Berghofer Medical Research Institute

agrf Empowering Australian Genomics

QFAB

qcif

Bronze partners

Pathtech First Choice In Science

customscience

华大基因 BGI

AB^{3}ACBS 2016 is hosted by:

QUT ife Institute for Future Environments

QUT ihbi Institute of Health and Biomedical Innovation